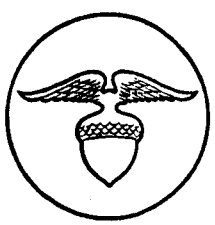
A PROGRAM OF RESEARCH DIRECTED TOWARD THE EFFICIENT AND ACCURATE MACHINE RECOGNITION OF HUMAN SPEECH

FINAL REPORT

prepared for

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION
ELECTRONICS RESEARCH CENTER
CAMBRIDGE, MASSACHUSETTS

CONTRACT NAS 12-129



Arthur D. Little, Inc.

A PROGRAM OF RESEARCH DIRECTED TOWARD THE EFFICIENT AND ACCURATE MACHINE RECOGNITION OF HUMAN SPEECH

FINAL REPORT

Prepared for

National Aeronautics and Space Administration
Electronics Research Center
Cambridge, Massachusetts

Contract NAS 12-129

Huseyin Yilmaz
Program Director

December 14, 1966
C-68366

Arthur D. Little, Inc.

TABLE OF CONTENTS

		Page
List of Figures		iii
I.	INTRODUCTION	1
II.	THEORY OF VOWEL PERCEPTION	11
	A. PHYSICS OF SPEECH SOUNDS	11
	B. PRODUCTION OF SPEECH SOUNDS	12
	C. PSYCHOPHYSICS OF SPEECH SOUNDS	14
	D. FUNCTION SPACE OF SPEECH	18
	E. TRANSFORMATIONS IN SPEECH SPACE	27
	F. TRANSFORMATIONS IN RUNNING SPEECH	29
	G. THE INHOMOGENEOUS TERMS	30
	H. VOICED VOWELS, TONE QUALITY	30
	I. TIME DEPENDENT CASE	32
III.	DESCRIPTION OF THE EQUIPMENT	34
	A. INPUT STAGE	36
	B. FILTER BANK	39
	C. LOGARITHMIC DETECTOR	40
	D. OUTPUT STAGE	40
IV.	EXPERIMENTAL SECTION	42
V.	DISCUSSION AND COMPARISON WITH OTHER APPROACHES	54
	A. FORMANT THEORY AND FORMANT TRACKING APPROACH	54
	B. VOCODER (VOICE CODER) APPROACH	55
	C. SEGMENTATION APPROACH IN SPEECH RECOGNITION	56
	D. SPEECH PRODUCTION, VOCAL TRACT, AND MOTOR APPROACHES	57
VI.	CONCLUSIONS AND POSSIBLE APPLICATIONS	59
VII.	REFERENCES	64

LIST OF FIGURES

Figure Number		Page
1	Energy spectra and format peaks of a voiced and whispered vowel. Vertical lines correspond to voiced case and are harmonics of the fundamental tone.	13
2	Under a given listening condition and for a given speaker a definite spectral (mel) distribution may be associated with each sound. This figure corresponds to vowel i.	19
3	Long-time average of speech sounds. Voiced speech curve is from the publications of the Bell Telephone Company. The whispered curve is obtained through the spectral ratios between voiced and whispered sounds.	21
4	Response functions of vowel perception are taken as Fourier functions. This is approximate in many ways but provides the vowel structure to sufficient accuracy. The effects of the rounding off on u_0 are shown on u_1 and u_2.	22
5	Sensitivity curve of the ear for general audio sounds and speech. Although sound is detectable from 16 to 20,000 c/s, only the 300-4000 c/s region contributes to speech intelligibility.	23
6	Formal representation of the speech space with Cartesian axes. The polar variables α_0, ϕ, and $\sigma = (\alpha_1^2 + \alpha_2^2)^{1/2}/\alpha_0$ are also shown. The latter corresponds to psychological attributes--loudness, chromaticity and saturation. In these representations a vowel is interpreted as a vector in a three-dimensional function space.	26
7	Vowel circle of human speech. This circle is a cross section of the speech cone at $\alpha_0 = 1$. The list includes the u sound (as in rue in French which is not used in English). We can see also that the sounds J and r are represented within vowel category, although we normally classify them among the fricatives.	28
8	A geometrical interpretation of the adaptation transformations--the transformation of the vowel vector to the reference frame of judgment. This leaves the relationships between speech sounds invariant.	31
9	Photograph showing the frequency analyzer and the vowel display device assembled as an integrated unit.	35

LIST OF FIGURES
(continued)

Figure Number		Page
10	Schematic diagram of frequency analyzer.	37
11	Schematic diagram of vowel display unit.	38
12	The analogue tristimulus functions $\bar{x}$, $\bar{y}$, $\bar{z}$ are obtained from u_0, u_1, u_2 by a linear transformation and represents the same theory. However, they are not orthogonal and therefore not convenient for identification of psychophysical attributes.	46
13	Running speech is passed through the filters shown. Note that these filters pass nothing of i and very little of α and u. Nevertheless, speech is perfectly intelligible and i, α, u are clearly heard. The experiment is analogous to the Land two-color projection and deals a heavy blow to formant theory.	50
14	The two-filter experiment is explained by the transformations. The space spanned by the two vectors f_1, f_2 is practically a plane. Yet after the transformations a circular (more properly elliptical) arrangement around the new α_0' axis is possible. Thus all vowel hues are obtainable by only two vowels. However, saturations will now depend on the vowel.	61

I. INTRODUCTION

The theory of speech perception introduced in this report falls within a more general approach to the problem of perception. This approach is evolutionary in its philosophy and statistical in its methods. The essence of the approach is as follows: first consider the physical properties of the stimulus energy and its statistical distributions in the environment; second, consider the needs of the organism in terms of individual and social survival. Given suitable neural material and biochemical processes, and given enough time for evolutionary forces to assert themselves, we then postulate that the perceptual devices evolved proceed toward a functional optimum. When supplemented with additional conditions of a constructional nature, and perhaps some restrictions related to early genetic fixation, the above statements are assumed to provide a suitable foundation to deduce mathematically the overall properties of a perception device.(1)

As in all evolutionary processes, the perceptual organization is a matter of compromise and balance between various stimuli in terms of their relevance and statistical distribution. The statistical attitude is here quite basic because perception devices are not designed for specific stimuli. Furthermore, we are not interested in specific designs or mechanisms but functional behavior of ensembles of devices under varying distributions of stimuli. This statistical character also removes possible teleological interpretations occasionally accorded to evolutionary arguments. Our theory does not say that such and such perceptions are needed and they must be evolved by the organism. This would be a very strong statement. We would rather consider the starting point of an organization as a chance mutation opening up the possibility of a class of perceptual devices under the constraints of the environment and potentialities of the organism. In fact, as new powers of perceptions become crystalized, new adaptations take place until, eventually, a fairly well-defined organization emerges from among the possible sets of devices. Neither is

the necessity to be mentioned ahead of time nor does the optimality need to be defined in an absolute sense. In other words, we are interested in a device which has a fairly stable behavior satisfying a condition of local optimum under constraints of partly environmental, partly genetic and habitual nature. Although the background is truly statistical, the language of probability need not always be imposed because the variables in question are also interpretable phenomenologically. In this way the theory can be presented similarly to thermodynamics as against statistical mechanics, and simple variational methods become applicable in terms of phenomenological variables. We shall have occasion to advance arguments of this nature in the next section. By and large the meaning of data and the identification of variables will be quite obvious.

Probably the simplest optimization process along these lines is to vary one or two parameters at a time. For example, to deduce the brightness sensitivity curve of the human eye one would assume the dimensions of retinal receptors and the lensed nature of the eye with its refractive index. Then under the known spectral composition of daylight illumination one would maximize the detection probability which is essentially the brightness curve. Of course one could then assume this curve as given and optimize for the size and shape of receptors, etc., eventually to come back to brightness for a second optimization. This procedure is mentioned not because it is done in actual calculations but because it is reminiscent of the self-consistent field methods of physics and suggestive of the actual way the perceptual organs might have evolved in nature.

The product of such an approach to perception is a functional theory rather than a particular mechanism. In general there will be many different mechanisms which can produce a given function. We shall consider the function as the more relevant question theoretically. Mechanism will be delegated to second place, and be considered as a representation of the requisite function. For example, the ear discriminates between the frequencies of sound to produce a psychological variable called pitch. This variable models the physical frequency continuum. For perceptual reasons

there is a need for the pitch to depend on the frequency as a function of a certain form. We shall see that an acceptable form is $p = C(\nu - \nu_o)^{\alpha}$ where C and α are constants. The discriminating mechanism may, however, be conceived in many ways. Two particularly familiar constructions imagined are cochlear resonator analysis (place theory) and neural peak detection (volley theory). Although, currently, controversy is going on between the adherents of the two theories, the distinction is unimportant perceptually. From our point of view one, or the other, or a combination of both is perfectly acceptable as long as the requisite function is producible within tolerable limits.

Incidently, from our point of view such phrases as "place theory" or "volley theory" are quite objectionable. These are not theories in the sense of well-formulated physical theories of the modern era, but both are only descriptions of a particular mechanism. For example, thermodynamics does not deal with a particular heat engine. Rather it is the theory of all possible heat engines and deals with thermodynamic variables independent of material and mechanism. Similarly, we are interested in an approach concerning classes of perception devices and their behaviors independent of their mechanismic content. If biology is to become a positive science like physics, we need also to raise the standards for what is to be called a theory. As in relativity, quantum mechanics, or thermodynamics we must strive for mutually consistent and logically complete sets of postulates from which experimental content can be derived in an unambiguous way. Note that in such a broad conception of methodology, theories which are usually considered distinct and attributed to different names may lose their identity and become different presentations of the same theory. For example, in color vision the Young-Helmholtz tristimulus theory and Herring opponents-colors theory are usually considered completely different, and in the past great controversy raged between the two. From our point of view one is obtainable from the other by a linear transformation of the basis vectors and, therefore, they both represent the same theory.[1] In physics a transformation of variables may carry us from Cartesian coordinates to polar coordinates but not into a new theory. Many of the

so-called new theories in biology and perception are of this nature and they stand in the way of further conceptual progress.

To be fair to the attitudes of the past we must remember that what was called a theory was actually the description of a mechanism. Thus Young-Helmholtz adherents assumed that the eye sensed the color qualities in terms of photosensors with certain absorption characteristics. Since the absorption curve cannot have negative parts, the Young-Helmholtz tristimulus curves, rather than Herring's opponents curves, are acceptable. On the other hand, Herring's followers felt that, psychologically, the brightness was a separate class by itself and red-green, blue-yellow qualities were mutually opposite sensations. They therefore ignored the tristimulus theory and sought to build a theory in terms of sensors of opposing nature. In the last few years it has become clear that even in the mechanismic sense the fight is quite unnecessary. G. Wald and his associates at Harvard have gone a long way to isolate, in the eye, photochemicals having Young-Helmholtz type of absorption characteristics. On the other hand, it was demonstrated that in the optical cortex of the monkey, opposing (+) and (-) electric potentials are produced by red and green stimulation of the eye. Thus the truth of the matter may be that the eye starts out according to a Young-Helmholtz mechanism at the retinal level; whereas, at the cortical level, the equivalent Herring representation is realized by a linear transformation.

Most of the current work in neurology and sensory perception today is directed toward the study of mechanism; the mechanism by which such and such effects are produced. This is of course a most useful activity; otherwise, how could we repair an organ when it is damaged or out of order? However, not all of us are doctors. (Doctors initiated biological research but their inherent interest is to cure the sick and the abnormal.) Some of us may want to know why the organ is there in the first place. Others might wonder if it could have been evolved differently and perhaps in a better way. Such questions lead us to consider a class of possible mechanisms which will perform a given task or function. The

pigeon and the man belong to different phila, but the brightness sensitivity function of their eyes is the same. The eye of the honey bee and the eye of a man are different in every conceivable respect, but both perceive colors in essentially the same way. The fact that they do and the question of why this is so is, no doubt, important to know.

As a result of long divergent evolution the compound eye of the insects and the lensed eye of the vertebrates are very different. In their separate ways these are the most complicated of all sense organs. The separation of ancestry of insects and vertebrates seems to have occurred about one billion years ago. The philogentic divergence of such magnitude would force us to conclude that practically no homological mechanism can exist between the two groups, and their perception devices must operate on entirely different principles. But then, why are the end results of color perception in bee and in man practically the same? We cannot understand this unless we assume that the task of perceiving colors is essentially independent of mechanism and that the two mechanisms are merely different ways of doing it. The only difference between the human color perceptions and bee color perceptions is that the bee spectrum is somewhat wider in range and the maximum sensitivity is shifted toward higher frequencies, from 555Å to 455Å. This shift is explainable from the fact that bees also see polarization. Indeed, as a consequence of Raileigh formula, there is more polarized light in the blue and ultraviolet region than in the red region, hence the shift as a result of optimization. Similarly, man's visible spectrum is narrower than the bee's because man's eye is lensed. For a lensed eye there is the additional problem of dispersion in focusing the image on a retina. This further restricts the spectrum.

Such considerations prompt us to consider a general set of questions. Given the external conditions of light and objects, what is the functional behavior of a visual perception device that will best help the survival of an organism? Clearly, the bee and the man will use different mechanisms to achieve such functions, as they use different mechanisms to breathe or to carry themselves from one place to another. But as long as

the visual conditions in the environment and the objects of interest for survival are similar, the functional behavior of their perceptions will be similar. Similar functions do not have to be performed by similar mechanisms. It follows that there are things which we can know without knowing these particular mechanisms. A knowledge of this sort is just as useful and important. For example, we can infer that the color vision of the bee and of the pigeon must be similar to that of the man without having studied the eye of the bee and of the pigeon. The same can be said for other insects and organisms. In this way we are able to conceive a theory of color vision which (with minor accountable variations) is the same for all organisms. This is indeed a valuable gain because, like the laws of thermodynamics, we have a theory of color perception independent of particular mechanisms or processes. Furthermore, we shall show in this report that the general laws of perception apply to the ear just as well as to the eye. In fact, it follows from the structure of the theory that there must exist a set of analogies between color and speech, and these can easily be tested by experiment. The general theory provides both the basis of this analogy and the extent to which it can exist. This knowledge of homology between the two senses can be turned into practical advantage in teaching the deaf to speak and the blind to read. Hope of finding this kind of sensory substitution possibility was from the beginning one of the motivations in our approach.(2) We think the functional sameness of color and speech leaves no doubt that the mechanismic approach is too primitive and narrow to provide a conceptual basis for psychophysics and perception. Here no analogy whatever exists in the mechanismic sense and yet the perceptual organization is the same as a result of the laws which are above and beyond mechanism.

In biological research we are obsessed with mechanisms. This is probably due to our desire to reduce everything to mechanics. Biology inherited from physics the mechanical point of view. It appears so natural to assume that when we understand exactly what is going on in the eye, we understand the eye. Yet the understanding of the human eye in this way would not tell anything about the bee's eye. It would not tell anything

about the deficiencies of either of them; if a favorable mutation occurred and some deviation from our cherished "norm" is observed, we would be trying to condemn it. We may come to marvel at some deficiencies or side effects which do not belong and which other organisms have long since eliminated. What is common between the human eye and the eye of the bee is not mechanism; it is not to be found in the composite or lensed structure, nor in the pulsed or continuous transmission of the messages, but outside the eye; namely, in the properties of light and the objects in the environment. In other words, the answer lies in the laws and regularities of the environment in which these organisms live and survive.

The ideal in biology--as in any other science--is to become like physics. But physics gave up the mechanical view a long time ago. The most important advances in physics since the turn of the century came through certain abstract, general principles with no mechanical background or ingredient. Quantum theory or the theory of relativity have no mechanisms. In fact, these theories can be said to be antimechanismic. No mechanical model can give light the properties it has, nor can any mechanism result in Planck's formula of quantization. The probability amplitude interpretation of the wave functions in quantum theory is impossible to explain by any mechanism. Yet the abstract principles of these theories are more powerful, more exacting and simpler than any mechanism that can be imagined. For example, when μ-mesons are absorbed in flight their lifetimes increase by the formula $t = t_o\sqrt{1 - v^2/c^2}$ where t_o is the lifetime at rest. A mechanist will try to find a mechanism which makes the lifetime increase by this ratio when μ-mesons are in motion. Imagine that he has found a mechanism which will do this for the μ-meson. But observation tells us that the lifetime of the uranium atom, the process which makes a dog age or a man think, is (as inferred from unquestionable indirect evidence) also slowed down by this same ratio when in motion by the same speed. Since the mechanism by which μ-meson disintegrate and the mechanism by which a man thinks cannot be the same, the mechanism of their slowing down must also be different. But then, why should all these mechanisms produce exactly the same ratio? The answer is that the effect

under consideration is independent and outside of mechanism. It is a structural property of the world we live in. It is a property of space-time in which all processes take place and all of them have simply to satisfy it. The value of this new statement is that without having the slightest idea of how a μ-meson distingetrates or why a dog ages we can make predictions about the behavior of their lifetimes when in motion. This is really a most powerful feature of relativity.

In a sense we are pursuing a similar theory in perception. Although we may not know the exact mechanism which produces color phenomena in bee's eye and in man's eye, we are able to predict the overall behavior functionally. This behavior is a property required by the laws and the regularities of the environment and the need to survive. Both organisms in their separate ways must meet the requisite function. Those species which did not evolve an efficient perception device must have vanished from the scene of struggle a long time ago.

Biology, we must always remember, is the science of living organisms that have survived. At the present stage of our knowledge almost anything that we may state about a biological mechanism will be most likely wrong or incomplete. Behind the human retina there are six layers of nervous interconnections of a most complicated nature. No one knows what they accomplish, let alone how they accomplish it. Undoubtedly, these structures are there for a reason, and what goes on in them is beyond present comprehension. Any conjecture about their construction or their working will be almost certainly erroneous or speculative. Of course, we would like to know what they do and how they do it. But for this we must wait fifty or perhaps a hundred years. But should we wait until we can follow every little ion or nervous pulse in order to understand color perception? We can see that the mechanismic approach can become rather fruitless with regard to certain questions which may otherwise be very simple to understand. To this end, however, we must pursue laws and statements which are independent of mechanism. One of the objectives of the present report is to demonstrate, with supporting evidence, the

practicality of such an approach for perception in general and for speech in particular.

The fundamental statements from which a coherent theory of perceptual organizations may be constructed are many in number, but they can be grouped under two principal categories:

A. Physics of the environment and of the carrier
B. Evolutionary history and the needs of the organism.

These are evidently very general principles. To be of direct use to our purpose we must make explicit statements within both categories. We shall only give statements which are related to speech perception and only those which have a direct relevance to the present task, that is, the vowel perception including a modest generalization to time-dependent speech sounds. In the first category we have:

1. Sound is the carrier of speech information.
2. Vocal tract modulation is the means of speech production.
3. Neural material poses no further restrictions.

In the second category we have the statements:

1. Perceptual organization models environment.
2. Perceptual variables optimize survival.
3. Percepts remain invariant under varying environmental conditions.

In addition to these, some further statements such as the Geshtalt properties of percepts, homomorphic rather than isomorphic correspondence of perceptual organization, etc., could be added. However, these will be largely unnecessary for the present task. The statements are actually a little stronger than need be. For example, instead of "optimize survival" and "remain invariant" we could say "tend to optimize survival" and "tend

to remain invariant." We hope the reader will be understanding about these rather categorical statements after we explain why we feel this is probably more appropriate. A theory is, in general, an abstraction delineating general structural properties of a subject. It is better to state the general structure as unambiguously as possible and study the physical situation relative to it; especially in evolutionary matters pertaining to perception this seems useful. We know that the human eye is not perfect enough to be called optimum. But if we have an idea of what an optimum eye looks like, we may then investigate the human eye as to the stage of its evolutionary development. For example, the human color vision has nonlinear properties such that as intensity is increased the red-green sensation saturates before the blue-yellow sensation. Clearly this is a deficiency but it gives us further evidence that the red-green process evolved later than the yellow-blue process and therefore is less securely established. The other evidence is that the red-green color blindness occurs much more frequently than yellow-blue blindness. Theoretically, our approach explains these as follows: the yellow-blue process is related to an eigenfunction which has two extrema and crosses the axis only once. The red-green function, however, has three extrema and crosses the axis two times. Consequently, the latter is expected to be implemented at a later stage of evolution.

II. THEORY OF VOWEL PERCEPTION

A. PHYSICS OF SPEECH SOUNDS

Sound waves being the carrier of speech information, we consider first the physical variables of sound distributions. Sound is a pressure wave of the air and, as such, it is a scalar function of space and time variables. A plane wave has a frequency ν, and amplitude A, and a phase ϕ. So far as the ear is concerned, it can be written $\Phi = Ae^{i\omega t + \phi}$. A complex sound can be considered as a linear superposition of such waves. In speech perception we shall often omit the phase variable. This is because the phase is not a determining factor in speech although it becomes relevant in other auditory phenomena such as binaural hearing, residue, and beats. This is quite understandable if we consider the nature of speech production as an activity of the oral tract. As stated in one of our postulates, the various parts and cavities of the oral tract modulate a carrier complex produced by the vocal chords. Speech information is therefore independent of the activity of the vocal chords which, in general, can produce phase-coherent sounds. Another and somewhat less relevant argument against the phase influence on speech is that speech sounds usually undergo many reflections from objects and are diffracted at corners and apertures. If phase were used as a perceptual determinant, intelligibility under these conditions would suffer drastic changes, which is undesirable from an evolutionary point of view.

Thus, as far as speech perception is concerned, we may discard the phase to a good approximation and concentrate only on the energy spectrum $I = |\phi(\nu, t)|^2$. An immediate consequence of this is that, theoretically, there will be no essential difference between the perception of voiced and whispered speech. Voice quality of speech, by and large, will be considered unessential to speech perception although, of course, very important for speaker identification and intonation. Furthermore, voice is much higher in energy than whisper and helps us to communicate at

considerable distances and in the presence of environmental noise. For simplicity, then, we shall begin the development of our speech theory by considering whispered sounds alone, which are modulated noise distributions. The resulting formulation should apply, with minor modifications, to voiced speech as well.

B. PRODUCTION OF SPEECH SOUNDS

In formulating a theory of speech perception it is important to enter into the subtleties of speech production much further than the above description. The familiar tone quality of the human voice is related to the fundamental and harmonic structure of the sound produced by the vocal chords. The output of the vocal chords is a series of puffs with the repetition frequency which is the fundamental. When these puffs are analyzed into Fourier series, the energy is seen to be distributed into a line spectra. In Figure 1 the harmonic content of a voiced vowel is shown. In the same figure the whispered form of the vowel is indicated by the noise distribution. We see that in both cases we are dealing with energy distributions over a range of frequencies. The fundamental frequency of voice which determines the voice pitch of the individual does not determine the distribution properties which carry the speech information.

A conspicuous feature in such distributions is that sound intensity is denser at certain areas than at others. The peaks are local energy concentrations called formants (Figure 1). They are related to the resonant characteristics of the vocal tract. For reasons of construction they are the sharpest concentrations obtainable by vocal tract modulation. Although in the traditional formant theory, a particular pair or triad of such energy peaks is supposed to determine a vowel, we shall reject their role in this sense and adopt the analysis of K. N. Stevens and A. S. House,[3] that they are simply the normal modes of the vocal tract. A consequence of this line of thinking would be that the formants correspond to some limiting situations of speech perception as spectral lines play, in color vision, the role of limiting saturations. In other

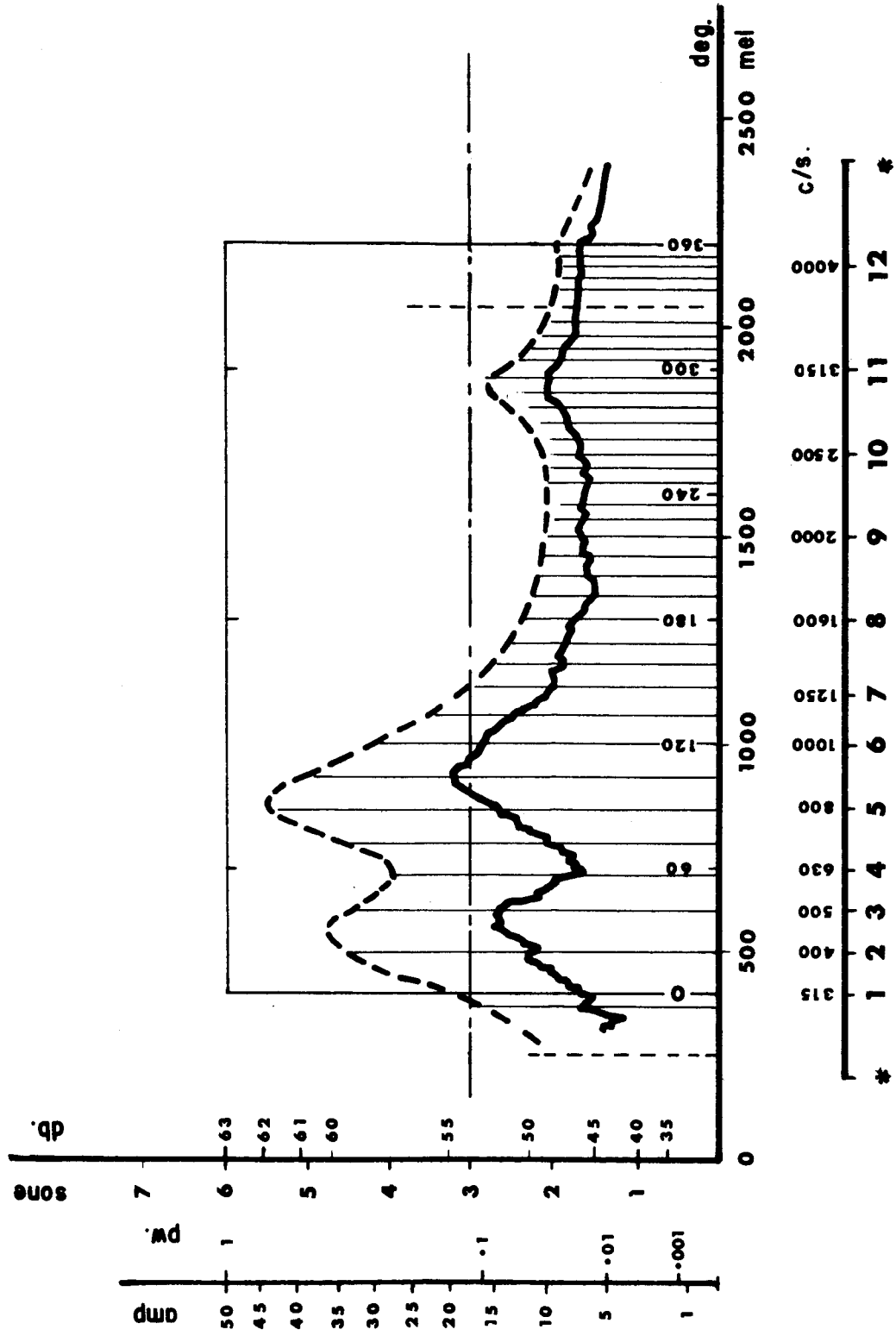


FIGURE 1 Energy spectra and formant peaks of a voiced and whispered vowel. Vertical lines correspond to voiced case and are harmonics of the fundamental tone.

words, formants would be a determining factor after they are properly weighed according to their position, strength, and sharpness but, in themselves, could not represent a unique theoretical determinant. In fact, a sharp enough formant would be something like a whistle, and we know that this limit is not a speech sound. It follows that a general theory of hearing which includes whistles would contain speech sounds, but the reverse is not necessarily true. In our present work whistles will be covered as a limiting case when a peak is very sharp. The normal formant peaks occuring in human speech are rather wide. Usually their width, that is, the width at 3 db below the maximum is about 10% - 15% of the center frequency. In terms of the mel scale, to be introduced shortly, they cover roughly 150 mels over the speech spectrum. Simply stated, these considerations imply that a speech theory can be constructed in terms of noise distributions alone, and the filters (detectors) need be no better in width than 10% of the center frequency.

C. PSYCHOPHYSICS OF SPEECH SOUNDS

According to our approach, the perceptual organization is modeled according to the environment but not necessarily in terms of direct physical variables. This modeling sometimes completely ignores a physical variable, sometimes transforms a variable into a psychophysical continuum to build a homological model of the external world. Thus we must assume that perceptual organization is to be considered in a psychophysical space but not in the space of physical variables. We here consider the problems of range, threshold, and psychophysics of speech sounds. Physically, the ideal threshold would be the thermal noise level below which the hearing sensitivity is absolutely useless. Speechwise, the threshold must be considered somewhat higher because the heart beats and breathing noises are also unimportant. Furthermore, there are, in the environment, noises, winds, and rumbles which are of the same magnitude or higher. Taking the speech threshold, $I_o(\nu)$, as the breathing noise in a quiet room (about 10 db above the absolute threshold) we next inquire into the psychophysics of loudness. As we know, the sensation of intensity does not grow pro-

portionately with the physical intensity. The form of the loudness function must satisfy certain general requirements of a perceptual nature. This requirement is the invariance of perception under the changes occurring in the speaking and environmental conditions. For example, when the overall intensity level of sound is increased or decreased, the speech must remain invariant. This means that if any two sensations, s_1 and s_2, corresponding to the intensities, i_1 and i_2, are compared, a common multiplicative factor K, as in Ki_1 and Ki_2, does not change the perceptual comparison. Thus we have

$$P(s_1, s_2) = f(i_1, i_2) = f(Ki_1, Ki_2) \tag{1}$$

This formula is rather general and contains both the logarithmic law of Weber-Fechner and the power law of S. S. Stevens.[4] Although acoustic equipment is usually made according to logarithmic scales (an indication that currently more people believe in the logarithmic law) and although in the task we consider the difference between the two is not crucial, we shall prefer the power law of Stevens. One reason, of course, is that it represents the experimental data better. Another reason is that it follows from our theoretical arguments directly, whereas the logarithmic law requires additional assumptions. To see this we apply our statement that perceptual variables model the environment. Thus to the variable i there should correspond a psychophysical variable s. Similarly, to the function $f(i_1, i_2)$ there should correspond the perceptual function $P(s_1, s_2)$, having a similar invariance property as above. From the invariance property, however, these functions can be written as $f(i_1/i_2)$, $P(s_1/s_2)$ and from the modeling postulate we have

$$P(s_1/s_2) = f(i_1/i_2) \tag{2}$$

Now let $i_2 = i$, $i_1 = i + \Delta i$. Expanding into Taylor's series and taking only the lowest two terms, we have

$$P(1) + P'(1)\,\frac{\Delta s}{s} = f(1) + f'(1)\,\frac{di}{i} \tag{3}$$

from which $P(1) = f(1)$, $P'(1)\frac{ds}{s} = f'(1)\frac{di}{s}$. The latter leads by integration in the vicinity of $i_1 = i_2$ to the power law of Stevens.

$$s = i^{\alpha}, \qquad \alpha = f'(1)/P'(1) \tag{4}$$

Experiment shows that in the dynamical range covering speech sounds we have very nearly $\alpha = 1/3$. In other words if acoustic intensity is increased 8 fold, loudness value s very nearly doubles. Loudness value is measured in sones and the scale is so chosen that an intensity of 40 db above the absolute threshold i_o is the unit sone.

The dynamic range of speech can be expressed in terms of sones instead of the usual db. Since speech is intelligible from the faintest whisper to the pain level of 100 db, the range is from 0.1 to 100 sones. Normal voiced speech at a distance of three feet creates a loudness $\sim$ 7 sones and a whisper at 10 cm (whispering into a microphone) corresponds to $\sim$ 5-6 sones. Note, in this connection, that loudness evaluation under the psychophysical function applies within a critical bandwidth which is roughly comparable to the center frequency. It two frequencies fall outside this separation, loudness adds independently. We believe this is significant in relation to tristimulus curves (to be introduced in Section V) which have approximately the same bandwidth. Without the above property these curves could not be considered linearly independent. Conversely, one may infer that the property emerges from a perceptual necessity in speech organization.

Next we inquire about the range and psychophysics of pitch sensation as related to speech perception. As we stated, the carrier of speech is the harmonic complex produced by the vocal chords. The fundamental and first few harmonics fall near the edge of the complex in frequency domain, and it is reasonable that lowest frequencies relevant to speech begin the second or third harmonics. Since the fundamental is of the order of 120 c/s for men and 200 c/s for women, a lower limit of 300 c/s appears reasonable. A limit much lower than this value would involve the fundamental

frequency of the vocal chords contrary to one of our postulates. The upper limit probably is set by the inability of the vocal tract to produce frequencies above 4000 c/s. Assuming that this is the case and knowing that (for other reasons such as the necessity of adaptation transformations) the delineation of the range is possible only within a considerable latitude, we shall take this range to be 300 to 4000 c/s or 250 to 3500 c/s, as the case may be. In other words, the speech sensitivity function, apart from possible transformations, will be considered to cover this range as a fairly smooth band-pass filter. The number of component filters covering the range may then be found, $n \simeq \log(4000/300)/\log(1 + 12.5\%) \simeq 12$. In our work we have used 12 to 14 commercially available filters.

As to the psychophsysics of pitch quality in this range, we may adopt again a power law interpretation of the standard pitch data. In the region of our interest this is given by

$$p = 35.2(\nu - \nu_o)^{\beta} \tag{3}$$

where $\beta = 1/2$, $\nu_o = 200$ c/s. This formula expresses pitch in the units of mel. The reason for the mathematical form is again understandable in view of the necessity of a perceptual transformation. From person to person the size of the vocal tract varies, with individual proportions remaining fairly constant. This shifts the overall frequency structure without altering the structure of speech. Thus, an adaptation transformation in the speech range can operate only if the pitch formula is invariant under a multiplicative transformation of frequencies, and the power law certainly has this property. In this way the original physical expression for intensity as a function of frequency, $I(\nu)$, is now to be interpreted as $f(p)$, where f and p are both psychophysical variables. These variables define a psychophysical function space in which the perceptual organization is to take place. Such an organization will be the subject of the next section.

We may note in passing that pitch function also has the property

of critical bandwidth. For example, the contribution of various frequency bands to speech intelligibility seems to follow a law; namely, the width appears to be about 150 mels over the whole speech spectra. This is roughly the half width of a typical formant. In other words, formants seem to correspond to certain linearly independent elements of speech perception, giving further support to K. N. Steven's interpretation that they are normal modes of the vocal tract. In fact the critical bandwidth for noise making is also of the same order. One would tend to conclude that these regularities arise from a perceptual necessity related to speech and its organization in the psychophysical space.

D. FUNCTION SPACE OF SPEECH

The aim of the perceptual organization in relation to speech is to recognize various speech sounds produced by different speakers through the spectral composition and intensity of these sounds. As stated above, the perceptual organization is assumed to take place in a psychophysical space, linear in mel and sones, and not in a physical space linear in frequency and intensity. Under a given listening condition and for a given speaker a fairly definite spectral (mel) distribution, f(p), can be associated with each produced sound (Figure 2). The problem of the ear is to recognize each relevant sound via its spectral composition. This means $I(\nu)$, or more properly f(p), may be expanded into some functional series

$$f(p) = \alpha_0 u_0(p) + \alpha_1 u_1(p) + \alpha_2 u_2(p) + \ldots \qquad (5)$$

where f(p) is any phonemic perception, and u_0, u_1, u_2, ... are linearly independent sensitvity functions. The coefficients α_0, α_1, α_2, ... determine the percept to within a transformation.

To find u_1, u_1, u_2, ... we resort to evolutionary and statistical arguments. If we consider u_0 the first sensitvity function to be evolved, we may well go back in time to the evolutionary stage before speech. As

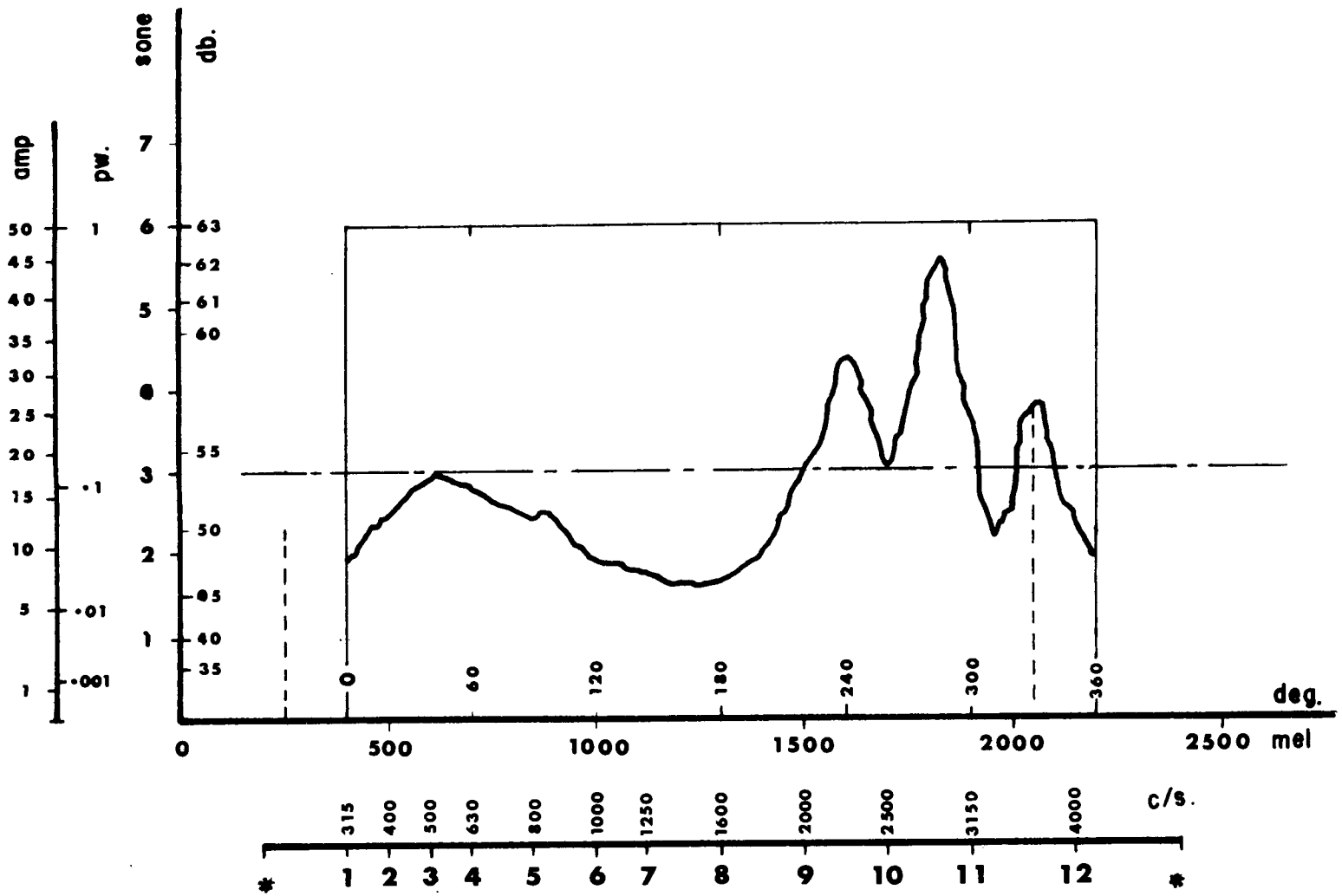


FIGURE 2 Under a given listening condition and for a given speaker a definite spectral (mel) distribution may be associated with each sound. This figure corresponds to vowel i.

in color then the u_0 can be assumed as the sensitivity to the presence of stimulus energy. Similar arguments of relevance and weight then lead to a range and a relatively flat audiocurve (Figure 3). In the case of speech the distribution is rather related to the long-time averages. Bell Telephone Laboratories published a long-time energy spectrum of voiced speech which we have also shown in Figure 3. This curve is for voiced speech. Since the voice part is not strongly correlated with speech information we need the long-time average of whispered speech. This is inferred by us from comparisons of voiced and whispered vowels of many people and by modifying accordingly the curve obtained by the Bell Telephone Laboratories. According to our previous arguments the end points of the curve at 300 c/s and 4000 c/s are rounded off as shown in Figure 3. If this is the relevant energy spectrum statistically, then u_0 must have the same shape in order to maximize sensations (Figure 5).

We consider u_0 normalized to unity (not a loss of generality) and inquire next about u_1. There is a perceptual requirement that u_1 be orthogonal to u_o so that percepts are separable, and that there is no overlap. The u_1 can be taken as the second member of a set of functions u_0, u_1, u_2, ... which are mutually orthogonal and each maximizes the response with respect to statitstical parameters in the stimulus distributions. For example, there is a class of stimuli which contains more energy in the long wavelength region than in the short wavelength region within speech spectra. Such sounds are u-like in vowel quality. If we subtract the overall average from the average of this class we find a function which is similar to a sine curve. The exact shape of this sine-like curve could be determined from the optimality of response and some mathematical conditions of regularity such as the continuity and boundary conditions. In the next step u_2 can also be determined from similar conditions, etc. Here instead of going through this process we simply resort to some well-known mathematical arguments: if u_0 was approximated as a perfectly flat function the end points of which go to zero sharply, then the set of eigenfunctions are the Fourier functions $u_0 = 1$, $u_1 = \sin\phi$, $u_2 = \cos\phi$. If, on the other hand, u_0 is a Gaussian shape, then the sets are the eigenfunctions of a

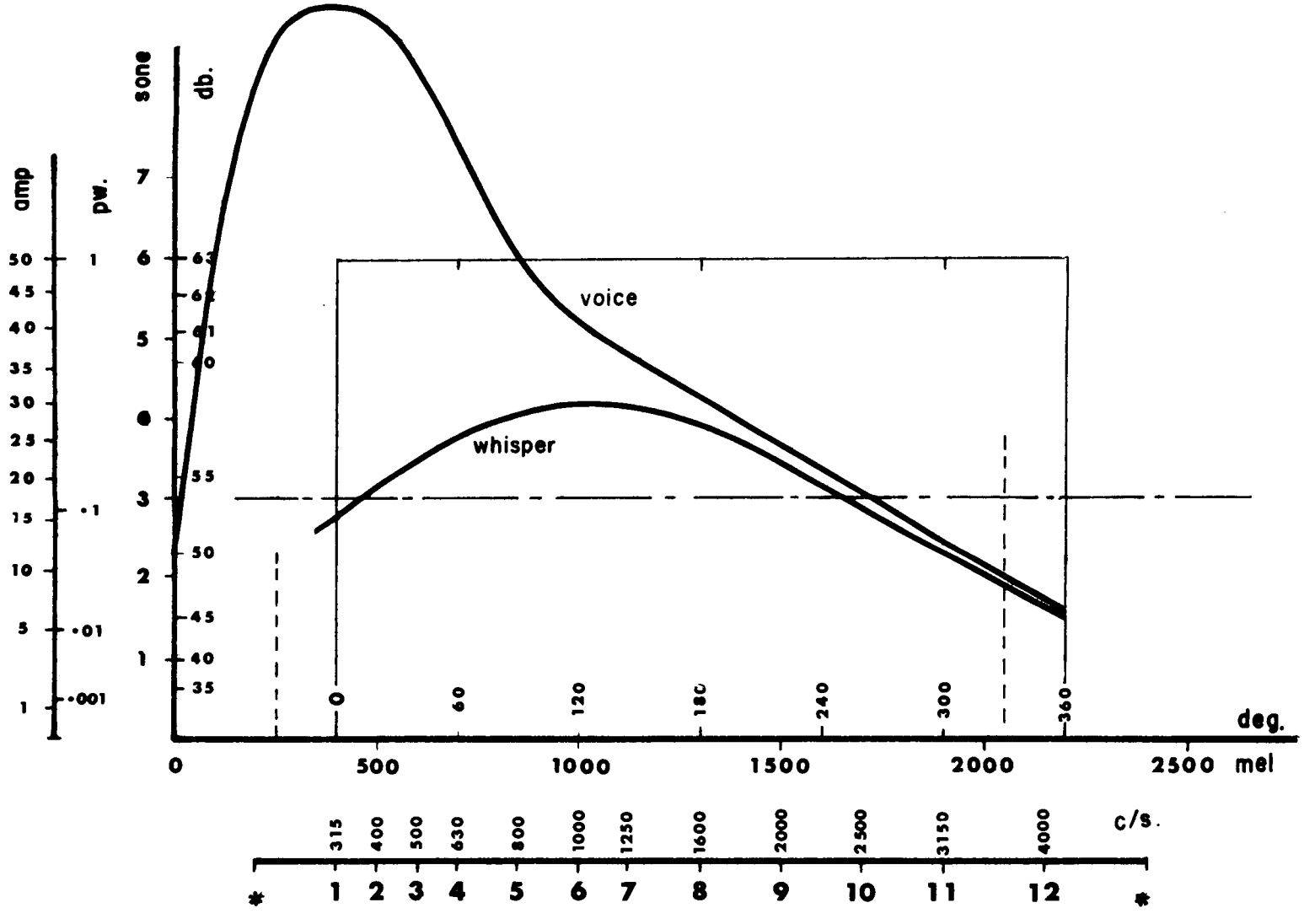


FIGURE 3 Long-time average of speech sounds. Voiced speech curve is from the publications of Bell Telephone Company. The whispered curve is obtained through the spectral ratios between voiced and whispered sounds.

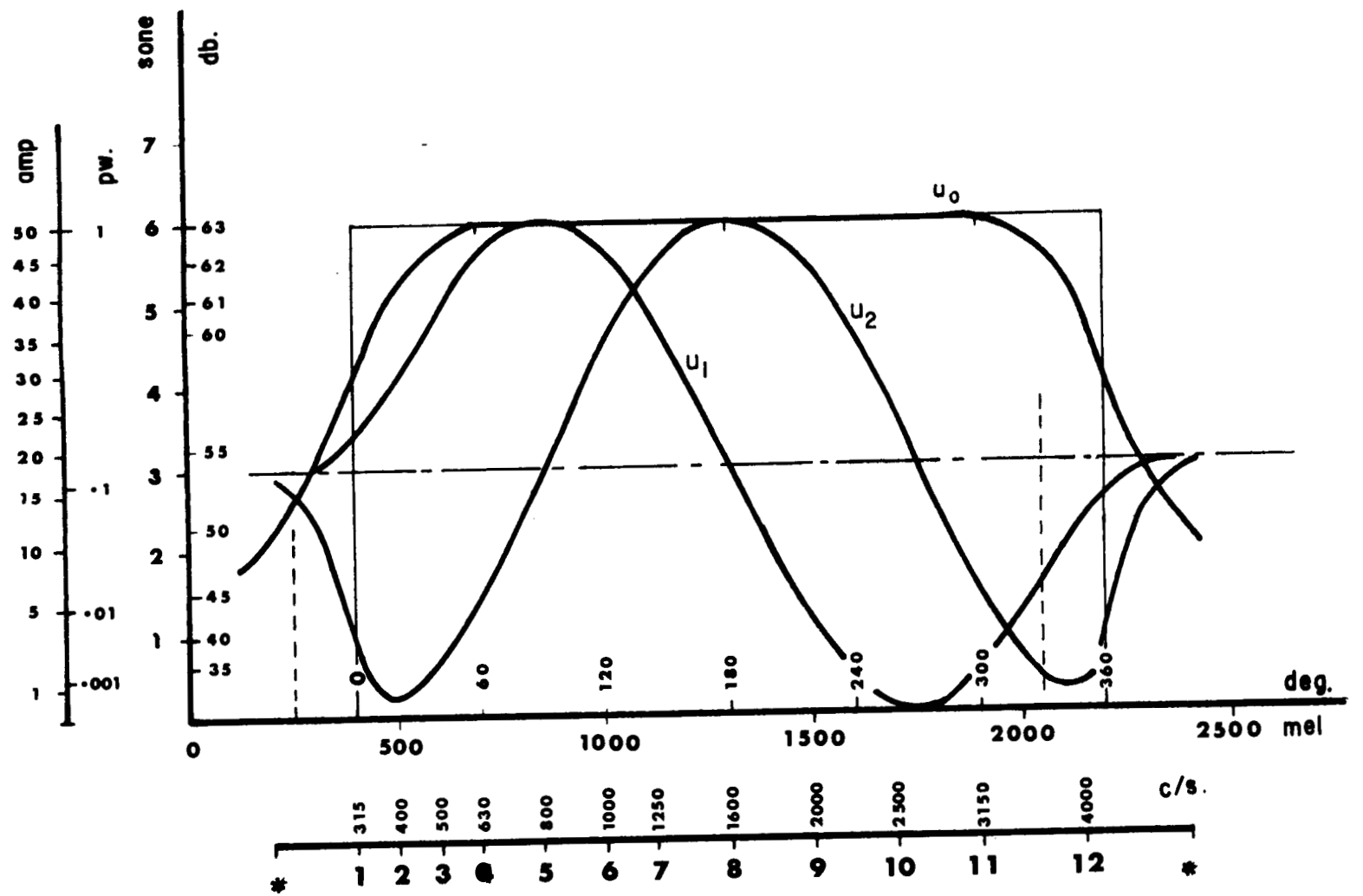


FIGURE 4 Response functions of vowel perception are taken as Fourier functions. This is approximate in many ways but provides the vowel structure to sufficient accuracy. The effects of the rounding off on u_0 are shown on u_1 and u_2.

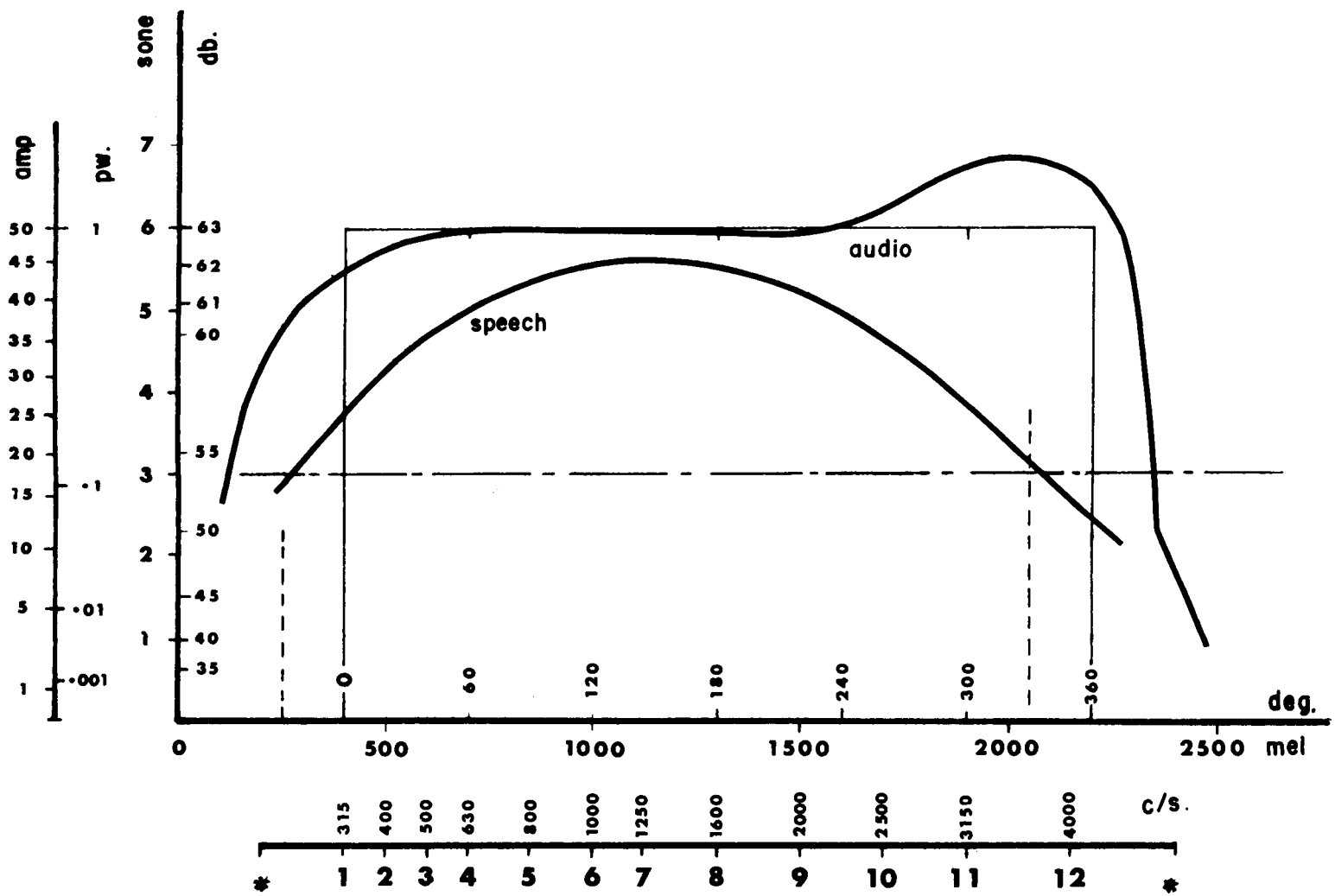


FIGURE 5 Sensitivity curve of the ear for general audio sounds and speech. Although sound is detectable from 16 to 20,000 c/s only the 300-4000 c/s region contributes to speech intelligibility.

linear oscillator $u_0 = e^{-p^2/2}$, $u_1 = pe^{-p^2/2}$, $u_2 = (1 - p^2)e^{-p^2/2}$. Although these are mathematically different expansions, they are fairly similar structurally.[1] We shall take the Fourier functions with minor rounding off of the eigenfunctions near the upper and lower limits (Figure 4). We believe that further refinements of the theory which adopt a slightly better representation will not alter the conceptual structure of the present work. We introduce, therefore, the angle variable of the Fourier series as a linear function of mel

$$\phi = 0.20\ (p - 400) \tag{6}$$

which gives ϕ in degrees (a mere relabeling of mel scale). It is interesting to note that arguments based on the sensitivity curves of the human ear would lead to similar conclusions. In Figure 5, the sensitivity data to audio frequencies and to speech intelligibility are given in mel and the sone scale. The audio curve is for 70 db and the speech curve is for 60 db.

The amplitude of Fourier functions (dynamic range) may be argued on two independent pieces of information. One of these is the actual vowel spectra in which low and high energy regions differ on the average by 20 - 30 db. The second clue is the related fact that noise will mask a given sound if noise is 20 db stronger than the sound. This means that the amplitude to which u_1 and u_2 are normalized will be 10 - 15 db. In the loudness region we have adopted, this is about 2 - 3 sones.

Regarding the number of eigenfunctions required, we argue as follows: The formant size covers approximately 150 mels, whereas the loudness summation is approximately three times larger. This means loudness summation extends 450 mels. This is just enough to allow sine and cosine curves, namely u_1 and u_2, but $u_3 = \sin 2\phi$ will result in perceptual aberrations. Although this is not sufficient reason to rule out u_3 and all higher functions, it is indicative at least that their weight must be considerably smaller. Another argument with regard to the number of functions

comes from the variability of formant positions for different people, and for one person under different articulations. This variability is often three times as large as the formant band itself. Consequently, the organization within which formants gain the meaning of perceptual elements must be about three times wider. This again leads to the estimate that perhaps the first three Fourier functions

$$u_0 = \frac{1}{\sqrt{2\pi}} \quad , \quad u_1 = \frac{1}{\sqrt{\pi}} \sin\phi \quad , \quad u_2 = \frac{1}{\sqrt{\pi}} \cos\phi \tag{7}$$

are essentially sufficient in speech perception. In any case, we shall adopt this minimum number and leave to the future any improvement to be brought about by the introduction of new functions. The three coefficients α_0, α_1, and α_2 of any distribution $f(\phi)$ are then given by

$$\alpha_0 = \int_0^{2\pi} d\phi \; u_0 f, \qquad \alpha_1 = \int_0^{2\pi} d\phi \; u_1 f, \qquad \alpha_2 = \int_0^{2\pi} d\phi \; u_2 f \tag{8}$$

Since u_0, u_1, and u_2 are orthogonal functions, a speech sound may be represented formally as a vector in a Cartesian system of three dimensions. The Cartesian coordinates of the vector are α_0, α_1, and α_2 (Figure 6). The same vector can also be represented by polar coordinates ϕ, $r = (\alpha_1^2 + \alpha_2^2)^{1/2}$ and α_0. These new coordinates have direct psychological meaning; namely, they determine the three psycholophysical attributes: loudness, vowel chromaticity, and saturation.

$$\alpha_0 \leftrightarrow \text{loudness}$$
$$\phi \leftrightarrow \text{chromaticity}$$
$$\sigma = r/\alpha_0 \leftrightarrow \text{saturation}$$

Note that loudness is a positive quantity indicating that all vowel vectors point into the upper part of space. Furthermore, there can be no α_1 and α_2 coordinates if loudness is zero. This means that all vowel vectors lie within a cone of maximum saturation (Figure 6).

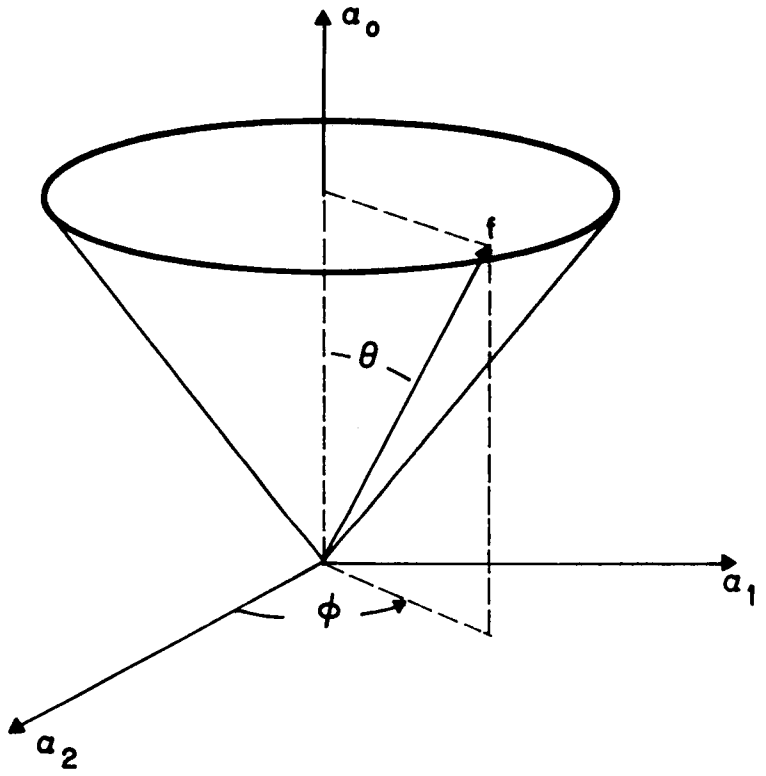


FIGURE 6 Formal representation of the speech space with Cartesian axes. The polar variables α_0 ϕ, and $\sigma = (\alpha_1^2 + \alpha_2^2)^{1/2}/\alpha_0$ are also shown. The latter corresponds to psychological attributes -- loudness, chromaticity and saturation. In these representations a vowel is interpreted as a vector in a three-dimensional function space.

$$\alpha_1^2 + \alpha_2^2 \leq \Sigma^2 \alpha_0^2 \tag{10}$$

Figure 7 shows a section of the speech cone with associated organization of vowels. In speech perception the maximum saturation cone may be taken as the loci of individual formants, although the loci of pure tones defines a larger cone with $\Sigma = 1$. It follows that pure tones carry a certain vowel quality, and the pitch must have a circular property. We must, however, refrain from further arguments in this respect and return to speech proper.

E. TRANSFORMATIONS IN SPEECH SPACE

Everyday conversation is carried out mostly under noisy listening conditions. The noises involved range from the rustle of leaves in the wind to the hum of the air conditioner in a room, and often possess vowel quality. Of course, these add physically to the speech spectrum and produce measurable changes in the sound distribution. In order to better serve the organism, the ear has evolved a perceptual transformation which removes the vowel quality from such sounds and reduces them to an achromatic noise quality. The purpose of the transformation is to leave speech content invariant. It can be considered as an adaptation phenomenon with a time constant larger than, or of the order of, that required for an average sentence (2 - 5 seconds). To a good approximation we shall assume this transformation to be linear.

$$\begin{aligned}
\alpha_1' &= \Omega_{11}\alpha_1 + \Omega_{12}\alpha_2 + \Omega_{10}\alpha_0 + \zeta_1 \\
\alpha_2' &= \Omega_{21}\alpha_1 + \Omega_{22}\alpha_2 + \Omega_{20}\alpha_0 + \zeta_2 \\
\alpha_0' &= \Omega_{01}\alpha_1 + \Omega_{02}\alpha_2 + \Omega_{00}\alpha_0 + \zeta_0
\end{aligned} \tag{11}$$

where primed α's are the new coordinates after the onset of, say, an air conditioner noise. The inhomogeneous terms ζ_1, ζ_2, and ζ_0 are small and

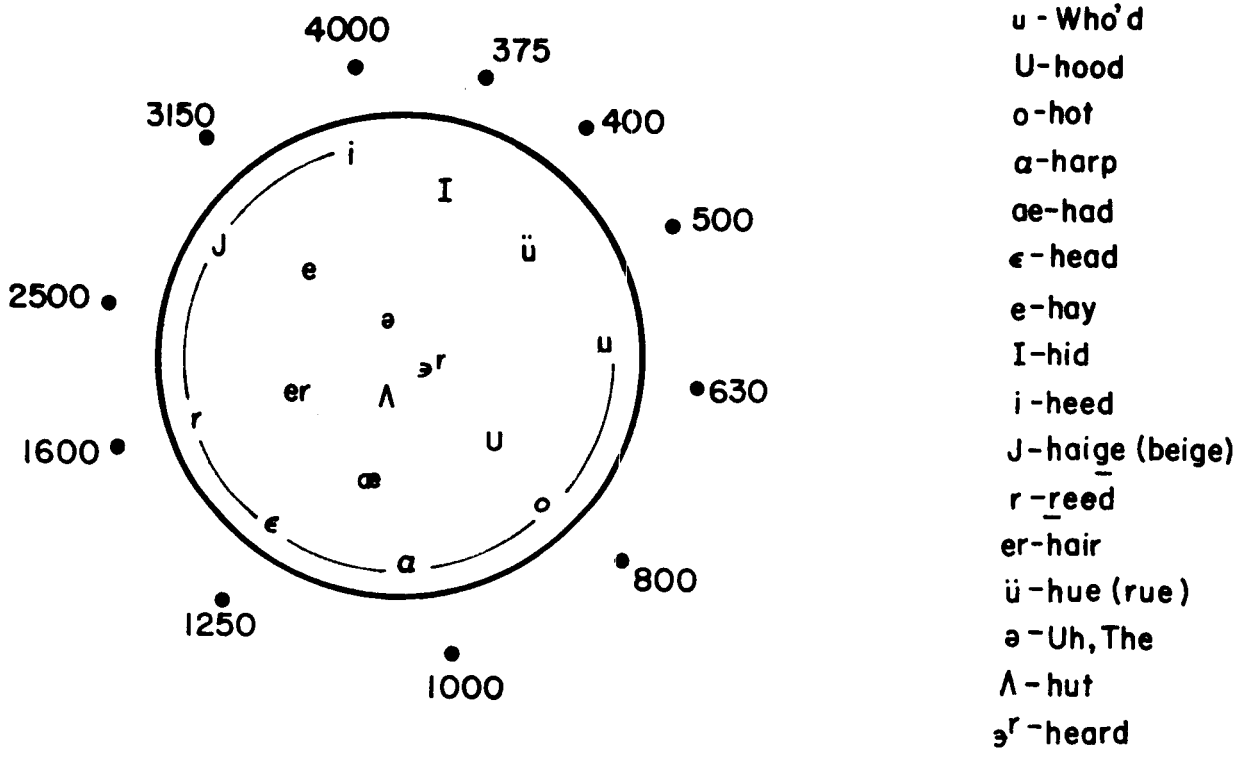


FIGURE 7 Vowel circle of human speech. This circle is a cross section of the speech cone at $\alpha_0 = 1$. The list includes the ü sound (as in rue in French which is not used in English). We can see also that the sounds J and r are represented within vowel category, although we normally classify them among the frictives.

usually negligible. To find the form of the transformation let the air conditioner possess a vowel quality lying in a α_1 direction, and let its saturation relative to previous adaptation be $\sigma = \alpha_1/\alpha_0$. Neglecting the inhomogeneous terms, we have, after the adaptation, $\alpha' = \Omega_{11}\alpha_1 + \Omega_{10}\alpha_0 = 0$ and $\alpha_0' = \Omega_{01}\alpha_1 + \Omega_{00}\alpha_0$ because the air conditioner does not have vowel quality in the new adaptation. This gives $\sigma = \Omega_{11}/\Omega_{10}$. From the reversibility of the situation, we find $\Omega_{10} = \Omega_{01}$ and from the solubility of the equations we get $\Omega_{00}\Omega_{11} - \Omega_{10}\Omega_{01} = 1$. An additional condition may be that the highly saturated vowels remain highly saturated after the transformation, namely, the maximum saturation cone is invariant under the transformation. With this fourth condition, all the coefficients are determinable, and we have

$$\alpha_1' = \frac{\alpha_1 - \sigma\alpha_0}{\sqrt{1 - \sigma^2}} \quad , \quad \alpha_0' = \frac{\alpha_0 - \sigma\alpha_1}{\sqrt{1 - \sigma^2}} \tag{12}$$

where the maximum saturation is taken to be unity ($\Sigma = 1$) as arranged earlier by a proper choice of units. We shall find that these formulae explain, to good approximations, transformation phenomena under steady noise. However, if σ is close to unity, these formulae will break down.

F. TRANSFORMATIONS IN RUNNING SPEECH

If the noise is not steady, or speech is running with a tonal quality, or with filter characteristic (such as listening in a room with bad acoustics), we need a weighting process to find the average saturation to be used in (12). This average will be defined over a finite adaptation time, T, and satisfy the conditions of achromaticity on the average,

$$\int_0^T \alpha_1'(t)dt = 0 \quad , \qquad \int_0^T \alpha_2'(t)dt = 0 \tag{13}$$

Evaluating σ with (12), we find the relative saturation to have the components,

$$\sigma_1 = \frac{\int_0^T \alpha_1(t)dt}{\int_0^T \alpha_0(t)dt} \quad , \quad \sigma_2 = \frac{\int_0^T \alpha_2(t)dt}{\int_0^T \alpha_0(t)dt} \qquad (14)$$

In these formulae, a weight function, say, of the form $e^{-k(t - t_o)}$ may improve the accuracy, but we did not have time to test such a possibility. However, the approximate validity of the above formulae are easily demonstrated.

G. THE INHOMOGENEOUS TERMS

The inhomogeneous terms ζ_1, ζ_2, and ζ_0 are investigated. They correspond to a perceptual necessity that when heard in the presence of environmental filtering actions, not only the vowel quality is added, but the magnitude of the vowel is altered according to its chromaticity. In Figure 8 one sees that in order to measure the angle and magnitude simultaneously in a new frame a shift of the origin of coordinates is necessary. As will be seen in the experiments with the Land analogue, the inhomogeneous terms lead to qualitatively unexpected consequences. This makes their explicit mention necessary here, although the magnitude of these terms is quite small.

H. VOICED VOWELS, TONE QUALITY

As argued earlier, the theory of speech we have presented is independent of voice quality. As the normalized curves of Figure 3 imply, our theory should apply to voiced speech as well if the lower part of the spectrum (the region 100 - 700 c/s area) is attenuated 15 to 20 db by a filter. In other words, apart from this readjustment, the expansion curves, transformations, vowel circle and all other concepts should, according to the theory, be universal and the same whether the speech is voiced, or whispered, or spoken by a male or a female.

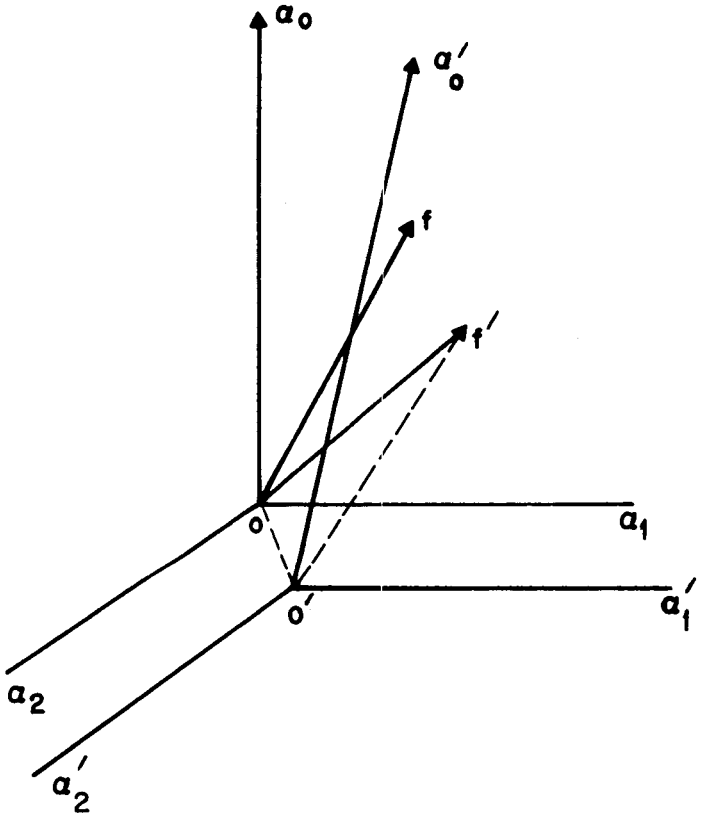


FIGURE 8 A geometrical interpretation of the adaption transformations - the transformation of the vowel vector to the reference frame of judgment. This leaves the relationships between speech sounds invariant.

The study of voice quality and the problem of speaker identification is a different subject than speech proper. We plan to study this in the future in relation to a theory of residue perception that we have developed.[5] At present we note only that all voiced vowels of a person may be constructed to good approximation from that person's achromatic vowel (that is, $\mathscr{r}$ sound) by multiplying it with standard (whispered) vowel curves.

Our theory of vowel perception is now conceptually complete, that is, mathematically self-contained within its assumptions. Minor readjustments on the range, threshold, even in the form of psychophysical functions, or the nature of expansion will not make qualitative change, although quantitatively better approximations would be achieved. Perhaps the implicitly assumed metricity of the vowel space is not valid to high approximations, but this will change only the second-order sequences of the transformations. First-order effects would still be true in a projective space where maximum saturation is not invariant. We shall not dwell at present on these finer points as we have not yet ascertained them.

I. TIME DEPENDENT CASE

Before we start the descriptions of equipment and procedures, it would be necessary to state that what we have so far is not a complete theory of speech, but only a vowel theory. Nevertheless, some of the experiments will have to do with running speech. For example, the transformation laws will be found valid for both vowels and consonants. We interpret this as an indication that the vowel theory really provides a basic framework for speech in general. If this is correct, a theory of consonants and of the running speech may be initiated on the basis of the following simple generalization

$$f(p, t) = \alpha_0(t)u_0(p) + \alpha_1(t)u_1(p) + \alpha_2(t)u_2(p) + \ldots \qquad (15)$$

where the time dependence of a general speech sound f(p, t) is separated into the α-coefficients. This would mean that speech is a time-dependent pattern in the speech space. The additional statements to cover this generalization and the necessary experimental procedures for their verification are being considered as the subject of the next phase of our work.

III. DESCRIPTION OF THE EQUIPMENT

The theory presented in the last section may appear rather abstract and elusive, but in essence it is sufficiently specific to provide all the required parameters and their numerical values for the construction of a device to test the validity and to demonstrate the practical usefulness of our approach. In this section we shall first describe the construction and performance of a device which we built on the basis of our approach. We shall then make suggestions for its further improvement.

The device shown in Figure 9 is a complete and selfcontained unit. It has its own power supply and a speaker. A four-track tape recorder is also incorporated as a sound source. Functionally, the device can be divided into two parts: (a) the frequency analyzer and (b) the vowel display unit.

The frequency analyzer is schematically shown in Figure 10. Input sound, such as flat noise or speech, is recorded on the two tracks, A and B on the tape. Track A is used as a reference track. The output of this track is fed directly through an amplifier and then to a speaker. Any sound we want to use as reference is recorded on track A. The output of track B is directed to a bank of twelve 1/3 octave Allison filters whose range covers 300 - 4000 c/s and whose widths are equal on the logarithmic frequency scale. The output of each filter is then directed to a potentiometer. To each potentiometer there is a dial attached so that the fraction of filter output taken is indicated by the dial setting. The output of the 12 filters is then summed by summing-amplifiers. Upon further amplification the summed output is fed to a speaker. A switch is provided to select either track A or tract B so that the filtered sound can be compared to the reference sound. The performance of such an analyzer has been satisfactory. For example, a vowel is synthesized by the device from flat noise through an appropriate combination of dial settings. A synthesized vowel is then compared with a natural

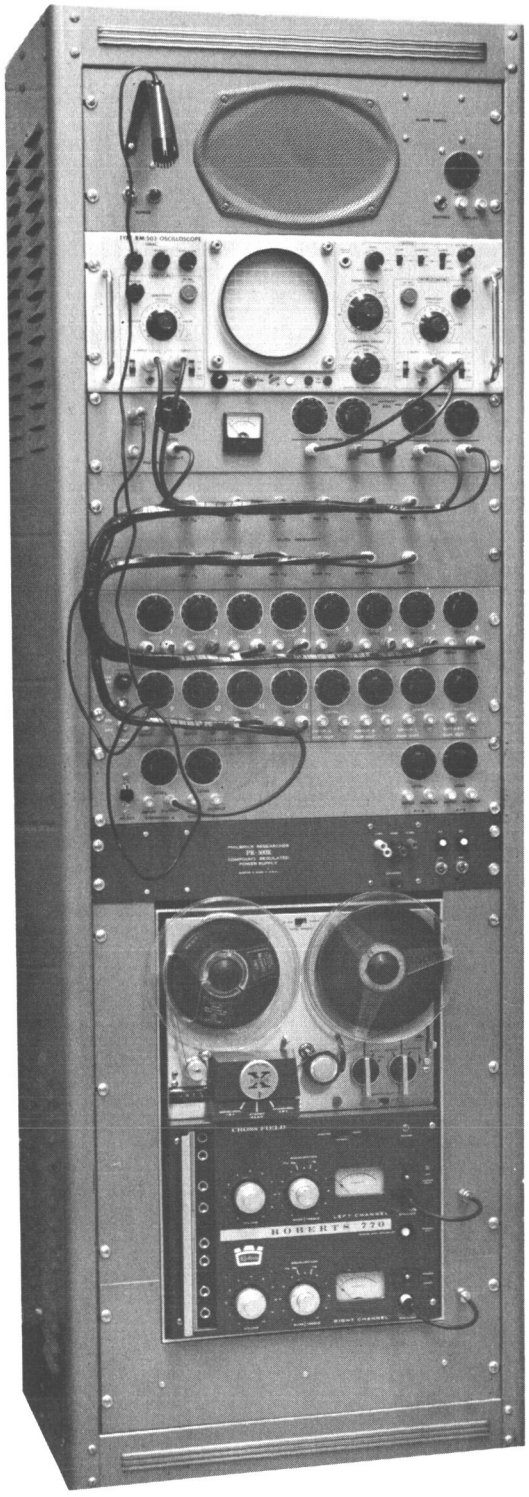


FIGURE 9

Photograph showing the frequency analyzer and the vowel display device assembled as an integrated unit.

vowel which has been recorded on track A and which the synthesized vowel is intended to be. The comparison is judged by ear. Adjustments are made to the dial settings until the synthesized vowel is perceptually close to the natural vowel. The dial settings on the analyzer then indicate directly the spectrum amplitude of the synthesized vowel. In this way the analyzer achieves the following: (a) the device renders quantities of measurements of synthesized vowel spectra which cannot be detected by the ear as different from the natural vowel spectrum and (b) the device can be used to demonstrate that all vowels can be formed from the neutral vowel (e) by just filtering with the simple functions provided by our theory.

From the theoretical discussion given in the earlier sections we see that the analyzing filters need only be as narrow as the formant width, that is, approximately 150 mels. Within the chosen range of 300 - 4000 c/s (1800 mels) this means $N \simeq 1800/150 \simeq 12$ filters, which is the number we have chosen for the filter bank in the analyzer shown in Figure 10. However, these 12 filters are equally spaced in logarithmic frequency scale as all commercially available filters are. This means that the width of filters at the high frequency end is approximately 200 mels, whereas those at the lower frequency are approximately 100 mels. Preferably, it is desirable to have filters which are of equal width in the mel scale. In order to achieve this a finer set of filters of width 1/6 octave would be more recommendable. Furthermore, the center frequency and the width of the band should be adjustable so that distortions caused by nonlinear or logarithmic amplification can be removed by adjustment. This is not possible for the present device.

The vowel display unit is shown schematically in Figure 11. It consists of the following parts:

A. INPUT STAGE

This consists of an impedance-matched preamplifier with built-in compensation for voiced vowels. Compensation is achieved by a R-C

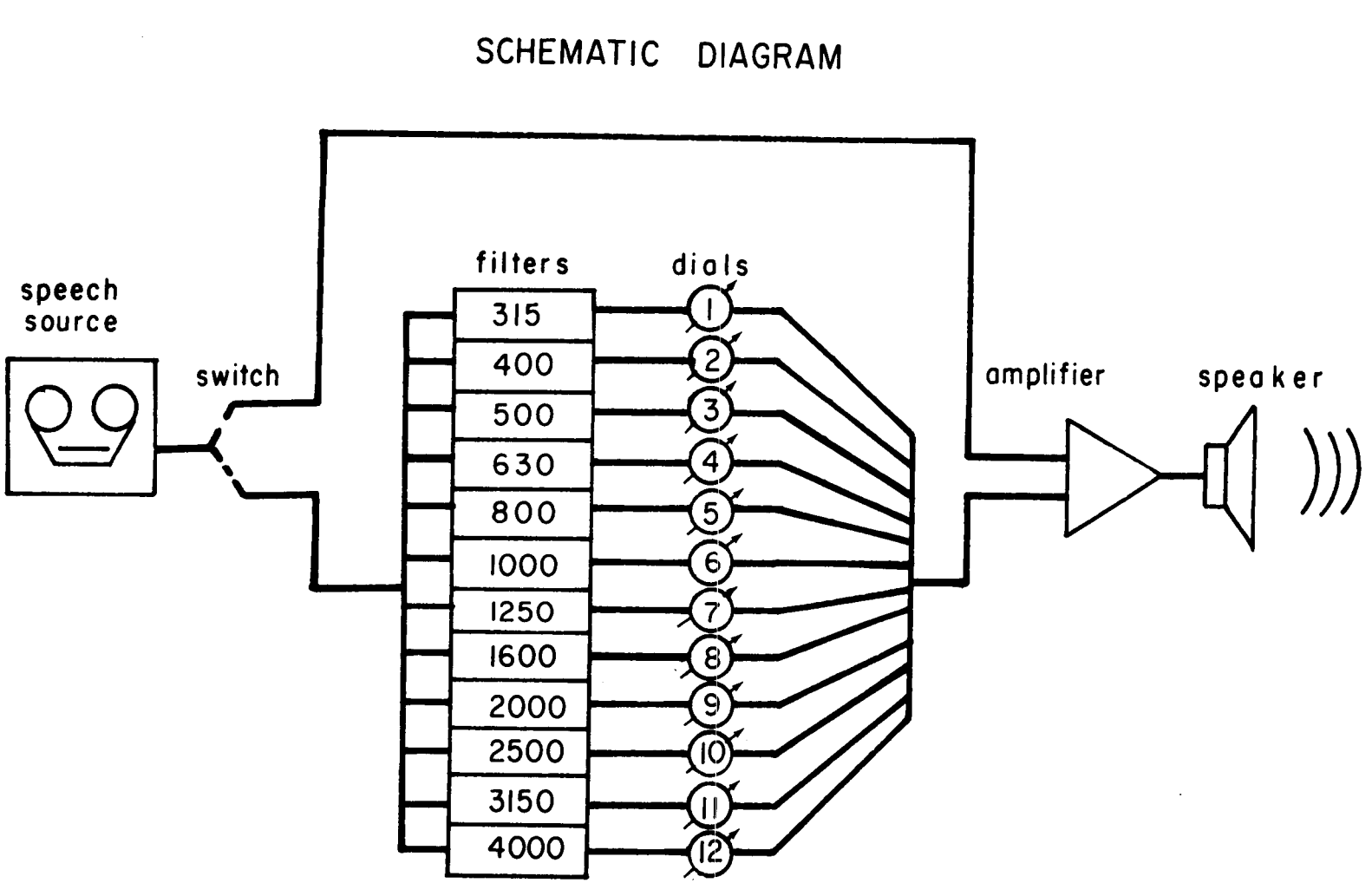


FIGURE 10 Schematic diagram of frequency analyzer.

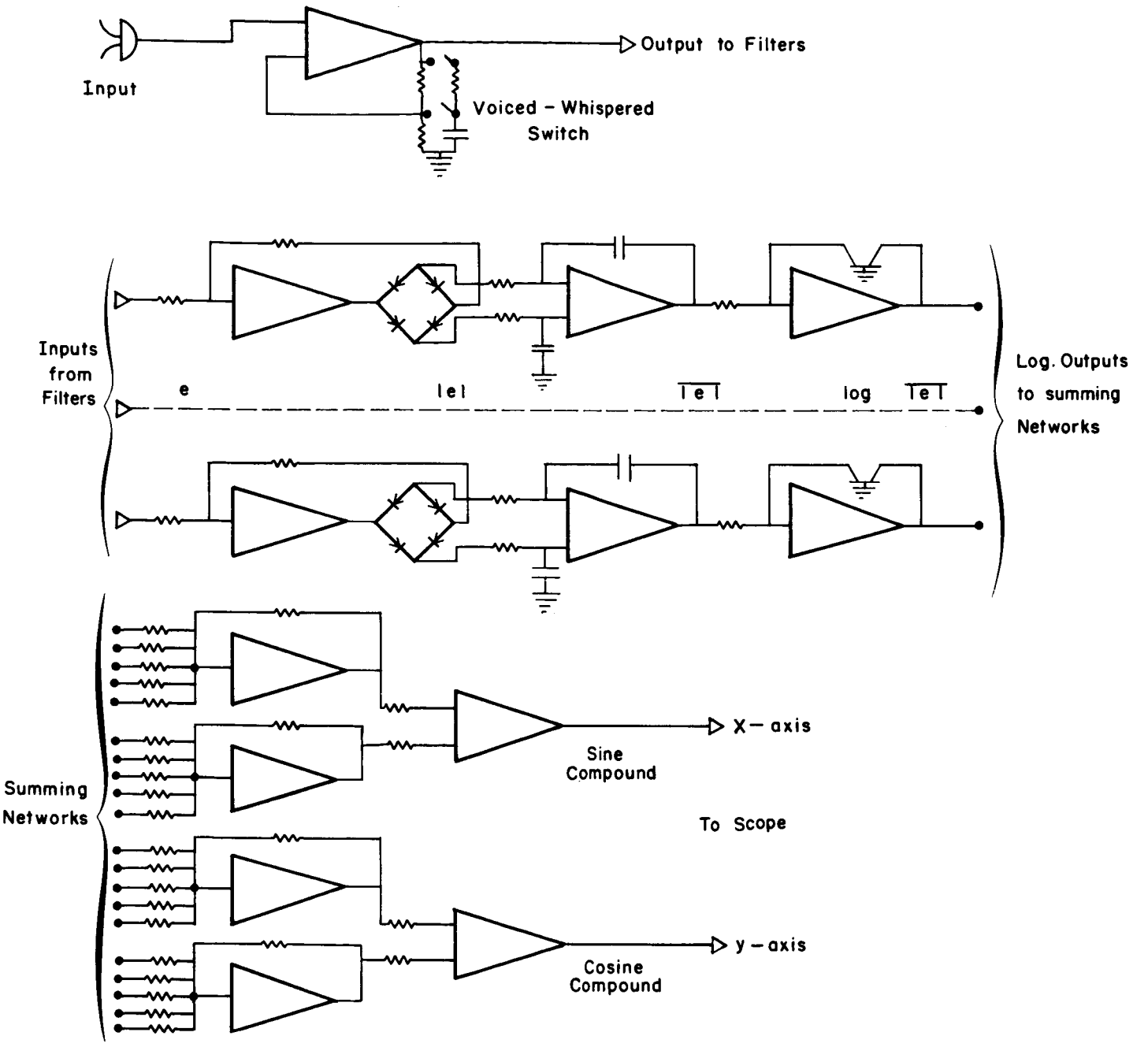


FIGURE 11 Schematic diagram of vowel display unit.

network, which provides an attenuation of 20 db at low frequencies, gradually decreasing to zero attenuation at high frequencies. A manual switch is provided which selects either the whispered channel or the voiced channel.

When the input sound is a whispered one, then the whispered channel is selected, and the compensation network is left out of the circuit. In this way the whispered sound passes the preamplifier without distortion. If the input sound is a voiced one, the voiced channel is selected. In this position the compensation network is incorporated, thus leveling off the build-up of the voiced sounds in the lower frequency range.

B. FILTER BANK

The filters in the bank are the ones used in the vowel analyzer described earlier. The equipment is so designed that the same filter bank is used both for the vowel analyzer and the vowel display device. The reason for doing so is to keep the cost of capital expenditure within limits. However, as we shall discuss later, a set of filters with characteristics different from those used in the analyzer is desirable for superior functioning of the vowel-display device.

Since the dynamic range of these filters is limited the amplification factor of the preamplifier must be made to vary so that its output level is always within its limits. A meter is incorporated to show the output level of the preamplifier before the signal is passed on to the filter. During each experiment the output level of the filter must be adjusted within the indicated range on the meter. As in the analyzer, the output of each filter can be varied so that it is possible in this display device to obtain any assumed sensitivity curve. The use of the sensitivity curve, as shown by our approach, is of extreme importance since it takes into account the perceptual properties of the ear.

C. LOGARITHMIC DETECTOR

It consists of a bank of 12 Philbrick logarithmic amplifier and diode rectifiers. With the aid of these amplifiers and rectifiers the output of the filters is individually detected, integrated, and logarithmically compressed to provide 12 signals, each proportional to db. These signals are separately weighed with resistors and then summed. The weighting functions that should be used have already been suggested by the theory discussed in the previous sections, namely, they are cosine and sine functions with rounded ends. In the present device the weighting functions are approximately by pure sine and cosine curves on the logarithmic frequency scale.

D. OUTPUT STAGE

The two final outputs, the sine and the cosine components of the logarithmic intensity spectrum after correction for the ear's sensitivity, are supplied to the X and the Y axes of an oscilloscope to provide the display. The Tektronix RM 503 oscilloscope is used. Any compensation of background noise and the centering of the (e) sound can be accomplished by the X-Y shifting of the oscillograph spot to the center of the scope, when such sounds are maintained at the input.

The vowel display unit described above gives satisfactory performance. However, for future construction of such display units we suggest that the following features be incorporated:

1. The filter bank should be composed of filters of width no wider than 1/6 octave or 50 mels. Each filter should be of equal width in the mel scale.

2. From the point of view of recognition, it is desirable to have each vowel displayed distinctly and separated from the rest. This can be achieved by slight modification of the weighting functions. Hence

it is advisable in future devices to use adjustable potentiometers rather than fixed resistors to produce the weighting functions.

3. R-C discharge networks of time constant 5 to 10 seconds should be added to the X-Y deflectors of the oscilloscope so that any persistent sound, such as background noise, is automatically discounted by the display device, thus making it more nearly a perceptual device.

IV. EXPERIMENTAL SECTION

As the construction of the device described in the previous section progressed, various experiments became possible to perform. In this section we shall present these experiments which we believe support the usefulness and essential correctness of the evolutionary theory. All the experiments reported here were suggested directly by the theory, and their outcomes were in agreement with those predicted by the theory. At times it has not been possible to carry the experiments to a high degree of accuracy but the quality and the magnitude of the effects have been ascertained beyond reasonable doubt. Further improvements on the quantitative aspects of the experiments seem to be merely a matter of time and facility. With better equipment and trained observers it should be possible to improve the accuracy and statistics. Well aware of our limitations of money and time we have not taken the direction of high refinement. There were so many predictions to be checked and each of them involved so much detail and ramification that we had time only to make sure if the effects were there perceptually, and if the direction and magnitudes were correct. Nevertheless, we have developed considerable confidence in the theory because we now have more than twenty experiments, any one of which could have given a result other than that predicted.

1. According to our theory there is a structural homology between the vowels perceived by the human ear and the colors seen by the human eye. The very first experiments presented are therefore the exact analogues of Newton's experiments on the nature of color. One of our immediate tasks has been the identification of a vowel which is the analogue of white. For reasons explained in Section II we defined the spectral composition of this "white vowel" as the long-time average of whispered speech.* Perceptually it is very close to the desaturated vowel phoneti-

*Due to the existence of transformations the precise definitions of a white vowel is not possible. For example, a continuous noise distribution is acceptable if its maximum and minimum differ from its average by less than 3 to 5 db. Such a distribution will sound as one of the desaturated vowels ɝ, ə, ʌ or something in between these (Figure 7).

cians designate ɝ (as in bird or birth with a slight tinge of German umlaute ü). For this identification we first compared the voiced and whispered vowels of various people and inferred the whisper average from the published Bell Telephone Laboratories' results on voiced average. Then this whisper average curve is constructed with the help of our filters by modifying the output of a standard "white noise" generator. When presented to the ear in suitable intervals (say, held 1/2 second every 2 seconds) it sounds like ɝ.

2. The whispered ɝ sounds of a person of either sex were filtered by our filter system to see if we could obtain all the other vowels of the same person. This was found to be possible in all cases. What was more, the filter settings which turned a person's ɝ into the same person's u, α, i, etc., showed some degree of universality, namely, the same settings would turn another person's ɝ into that person's u, α, i. This statement is also essentially true for voiced vowels and lends definite support to the correctness of the present approach. It corresponds to Newton's prism experiments in color.

3. Conversely, the u, α, i, e, etc., vowels obtained by analyzing a person's ɝ were summed together (graphically) and the produced resultant sound was perceived as the original ɝ sound. This is analogous to Newton's synthesis of white by passing colored light produced by a prism through a second prism. It may be said that it is also the analogue of Maxwell's production of white by rotating a top with colored stripes on its periphery. Another experiment of the same nature would be to sum all the vowels produced by a person and listen to the resultant sound. For voiced sound this is quite difficult because one must ascertain that all the vowels are produced with the same fundamental frequency. This has not been possible for us to do. For whispered vowels it is easier. We have summed the vowels of a person graphically and reproduced the same resultant shape by filtering white noise. It was indeed essentially the same as the whispered ɝ.

4. As in color there exist vowel triads which may be called "primaries." The meaning of these primaries is that by a suitable combination of three primaries any other vowel can be obtained. Although the choice is not unique, and many triads can be chosen equally well, the saturated u, ε, and i shown in Figure 7 forms a particularly convenient set to demonstrate the primaries. All vowels obtained from these primaries were sufficiently acceptable. The sum of i and ε gave e, the sum of and u gave α, o, u, etc. The produced vowels were not always as saturated as the primaries (also true in color) but they were acceptable as clear vowels for all practical purposes.

5. If two whispered vowels were mixed and then presented through the same loudspeaker, the sensation was a sound which could be represented on the vowel circle as a point lying between these two vowels (Figure 7). This is in agreement with our prediction that the point will lie at the centroid of the two vowels. This statement shows that laws similar to those of Grassman are also operative in vowel space. The only exception here is that although in color space pure frequencies are still colors, in audioperception pure frequencies have qualitatively a further whistle-like distinction. Two pure frequencies do not add up to a third synthetic sensation but are separately perceivable. This statement is not true for whispered vowels which have a width of the order of 150 mels or more. Thus as far as vowels are concerned the ear behaves synthetically like the eye. It does not behave analytically as is often held by students of audioperception. Figure 7 shows in a somewhat exaggerated way the delineation of the domain of whistles and vowels.

6. The question of vowel quality perceived in pure frequencies is an interesting one. We have made some effort in this direction but we were subsequently informed that G. Fant has made extensive research on this subject.[6] The result of his work is consistent with our diagram (Figure 7), that is the 400 - 650 c/s range is associated with u, the 650 - 900 c/s range with o, and the 900 - 1100 range with α, etc. Our own experiments and

the work of T. Chiba and M. Kajiyama support this conclusion.[5] Thus the theory automatically provides a reasonable explanation of these data pertaining to pure tones, because pure tones represent in our theory the limiting case of maximum saturations.

7. From a series of experiments the analogues of C.I.E. tristimulus curves were inferred. The method we used is again similar to that in color and involves the prepresentation of the data on vowel mixtures by three non-negative smoothly varying functions. Since we are interested in vowel quality alone, the problem is two dimensional and one of the functions can be chosen within quite a latitude. The problem is then to represent various distributions heard by the human ear in such a way that the formulae

$$x = \int \bar{x}(p)f(p)dp \quad , \quad y = \int \bar{y}(p)f(p)dp \quad , \quad z = \int \bar{z}(p)f(p)dp$$

provide the correct coordinates of mixed sounds. Here f(p) is the mixed distributions and $\bar{s}(p)$, $\bar{y}(p)$, and $\bar{z}(p)$ are the tristimulus curves. This process turned out to be much more involved than we originally assumed. What was done instead is to infer the adequacy, within present experimental accuracy, of a set of three theoretically deduced curves (Figure 12). These curves correspond roughly to u, ε, and i sounds and are in essence a linear transformation of the Fourier functions u_0, u_1, and u_2.

$$\bar{x} = u_0 - \frac{1}{\sqrt{2}}\left(\frac{1}{2}u_2 - \frac{\sqrt{3}}{2}u_1\right)$$

$$\bar{y} = u_0 + \frac{1}{\sqrt{2}}u_2$$

$$\bar{z} = u_0 - \frac{1}{\sqrt{2}}\left(\frac{1}{2}u_2 + \frac{\sqrt{3}}{2}u_1\right)$$

We can see that the new functions are no longer orthogonal. An interesting fact with these functions is that if we insist on two of them having only one maximum each, then the third function must have two maxima. This is, of course, true also for color vision. A long time ago, when this was

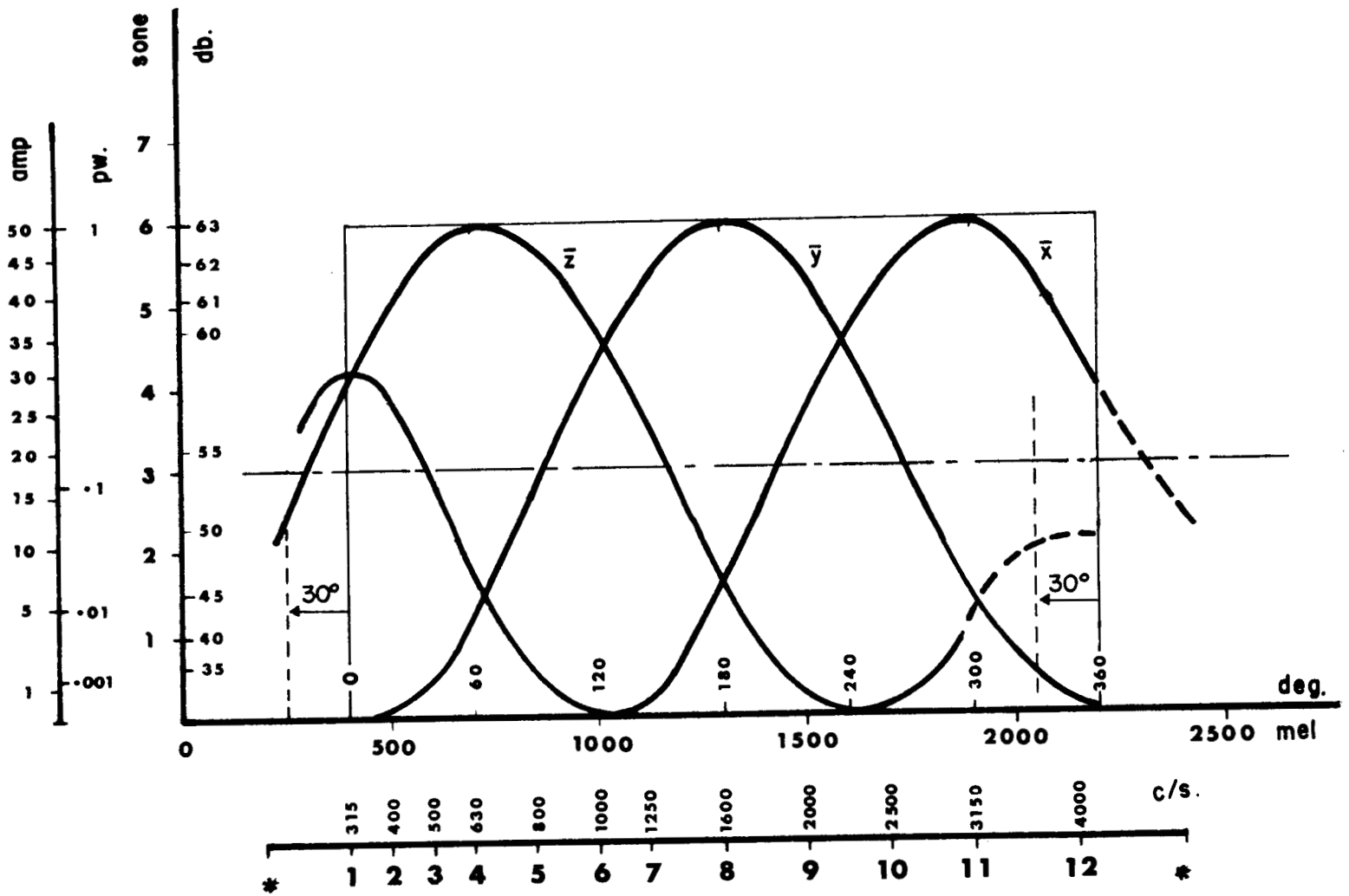


FIGURE 12 The analogue tristimulus functions $\bar{x}$, $\bar{y}$, $\bar{z}$ are obtained from u_0, u_1, u_2 by a linear transformation and represent the same theory. However, they are not orthogonal and therefore not convenient for identification of psychophysical attributes.

experimentally discovered in color vision it was thought to be a coincidence. The argument went as follows: The photosensors of color work on the basis of absorption of light and the sensitivity functions are absorption curves of certain chemicals such as rodopsin. By an evolutionary quirk the third photochemical happened to have a second bump. We can see from our theory that this is no quirk at all and appears necessary perceptually for the circular arrangement of colors and vowels. For example, the familiar (i) sound of human speech is perceived with the help of this second maxima.

8. In the case of voiced vowels the following experiment is interesting. When one pronounced a vowel, say α, on various keys (by changing the fundamental frequency), the position on the vowel circle did not change. This fact was directly visible on the oscilloscope. In other words, the theory is independent of voice frequency and the device extracted only the vowel quality. Similarly, whether a vowel was pronounced by a man, woman, or child the oscilloscope extracted and displayed the vowel correctly.

In another series of experiments speech analogues of color transformations are investigated. It is known that the human eye can adapt itself to different illuminants. For example, objects preserve their color appearance under direct sunlight illumination, under blue sky illumination (when the sun is covered by clouds), and even under artificial light. Such adaptation is accomplished by transformations in the eye. In this process the eye transforms away the illuminant quality and restores relative order of object colors. To establish such transformations in vowels we did the following experiments.

9. Any whispered vowel when continually heard degenerated perceptually into white noise. This was tested by recording a vowel into a loop and playing it continually. After a time passage of 3 to 5 seconds the sound became no longer a recognizable vowel. This is analogous to the fact that under sustained illumination by a colored light, a room still appears to be achromatic.

10. Any whispered vowel, other than ɣ , sounded more distinct when introduced under the background of a desaturated vowel. This phenomenon is analogous to contrast phenomenon in color.

11. When speech was passed through a filter representing a desaturated vowel quality, such as (a), the speech remained intelligible although physically the (a) ingredient was predominant. The ear discounted such permanent ingredients as noise and restored the relative relation of speech sounds. This phenomenon is analogous to wearing tinted sunglasses through which a permanent tint is produced but the eye restores the color of the object in its original order. Note that without the idea of transformation this experiment would have been a contradiction because according to the laws of mixing every time the speaker pronounced (u) the ear is supposed to hear (o).

12. In color there exists an after-image phenomenon. For example, if one stares at a piece of red paper for a long time and then quickly shifts the eyes to a sheet of white paper, the white paper would first appear as bluish-green (complement of red). The analogy of this phenomenon was also observed in vowels. We listened for a long time to a whispered (u) and then switched to ɣ ; the perception of the latter was found to be (ae)-like. This experiment was performed on other vowels. The same effect was observed. These effects were tested only on a few people. In the future we intend to undertake the task of testing these phenomena on a large number of subjects and obtain statistically meaningful averages.

We have also investigated the vowel analogue of Land phenomenon in color. The existence of such an analogue is a consequence of our theory of vowels. We can state now that this predicted analogue is also established by our experiments which are described below.

13. Natural speech after passing through a narrow band-pass filter (of width $\pm$ 5% of center frequency) became virtually unintelligible. In particular, vowel qualities were no longer discernible. This is anal-

ogous to the fact that in color vision the eye will not make adaptations to colored glasses if the saturation is less than 8% (Helson 1938).

14. Natural speech after passing through two such filters, suitably separated in frequency, become intelligible to a considerable degree. The intelligibility was found to vary depending on the frequency of the band-pass filters used. The pair of filters used were those with frequencies centered in the (u) and (a) regions (Figure 13). This is completely analogous to the Land experiment in which all colors were obtainable by two sharp filters.

15. Instead of using natural running speech, separately pronounced vowels (u), (a), (o), and (i) were passed through the same pair of filters. These vowels were found to sound like a distorted (o), and the vowel (i) virtually disappeared. This is analogous to the fact that in Land projection if one uses large single-color scenes instead of a natural scene, the observed colors will all tend to an orangeish wash, and blues and violets will be missing. However, if the vowels are produced in a connected way as in natural speech, then all vowels including (i) are clearly perceived by the ear. This phenomenon is especially remarkable because there is no frequency component and no formant in the frequency range which corresponds to the vowel (i). The fact that (i) is physically taken out is made visible to the eye both by the knob settings and by the oscilloscope display of the knob settings and by the oscilloscope display of the device. It is a fascinating fact of the human perceptual device. Again, its Land analogy is that blues and violets are observed although no spectral components corresponding to these colors existed in the projected scene.

16. In all these experiments the intelligibility of the speech and the discernibility of vowels were initially very poor but improved with time. This is analogous to the case in color and is interpreted as a time-dependent adaptation transformation. After full adaptation no further improvement occurred. A similar effect was observed previously

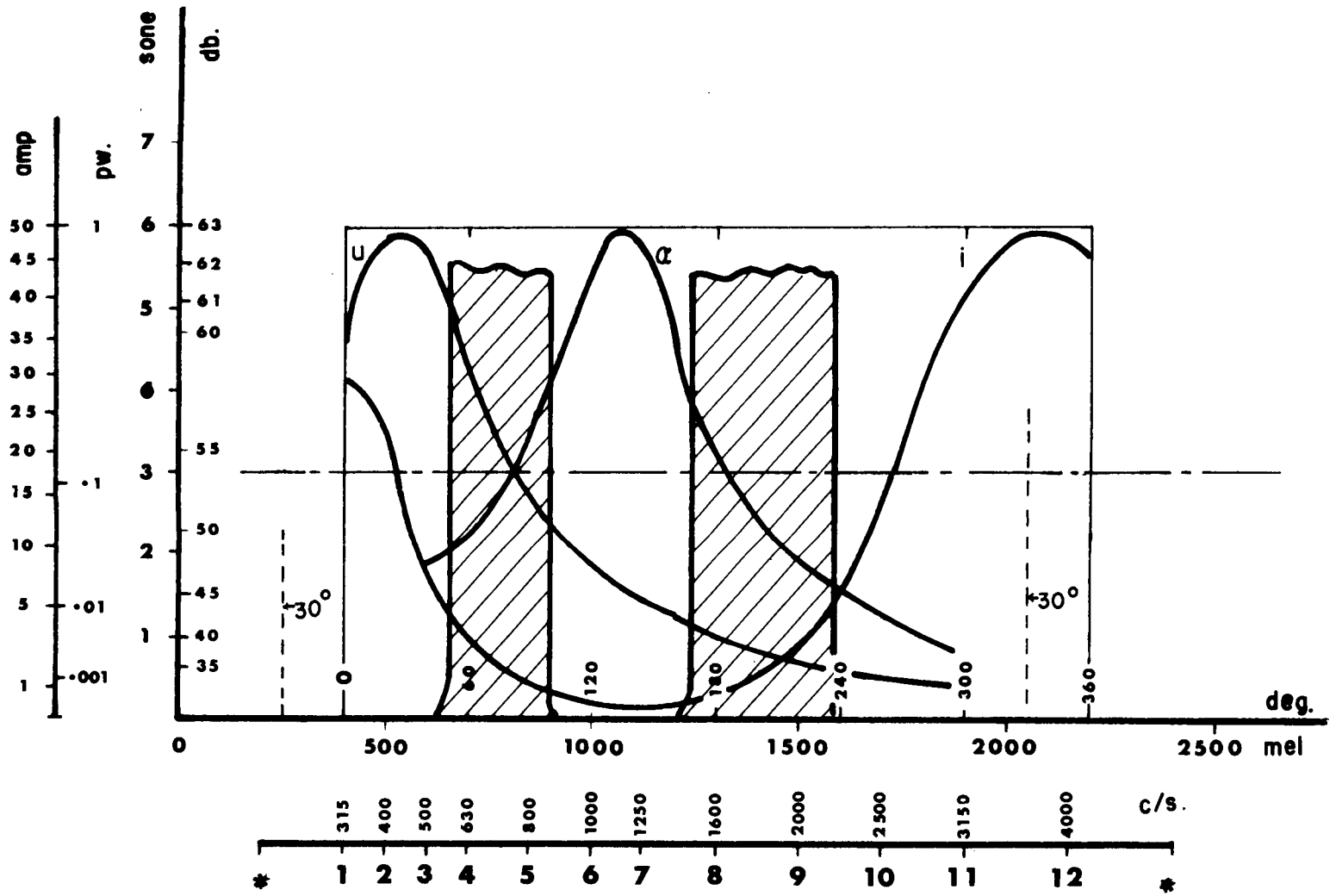


FIGURE 13 Running speech is passed through the filters shown. Note that these filters pass nothing of i and very little of α and u. Nevertheless, speech is perfectly intelligible and i, α, u are clearly heard. The experiment is analogous to the Land two-color projection and deals a heavy blow to formant theory.

by Chiba and Kajiyama.[6] They interpreted this effect as though increased attention and familiarity caused the increase in intelligibility. However, intelligibility is again reduced at the start. This is true no matter how familiar or how much concentration a person gives, proving that the phenomenon is actually an adaptation transformation similar to that in color.

17. To test the question of vowel complementaries we passed white noise through our bank of filters and investigated which pairs of filters, when simultaneously turned up, contributed only to an increase of loudness without influencing the vowel quality. These turned out to be approximately the filters with center frequencies (in c/s) 400 and 1250; 500 and 1600; 630 and 2000; and 800 and 2500. Note that any one of these filters when considered singly adds vowel quality; for example, when 400 is turned up it contributes (u), when 1250 is turned up it contributes (a), when 400 is turned down it contributes (a), etc. Turning complementary pairs up or down simultaneously seemed to contribute, apart from intensity change, a slight perceptible effect of harshness and softness, respectively; but this was clearly a small effect and in any case was not a vowel quality.

18. The preceding paragraph could be taken as a proof of the non-existnece of fourth and fifth terms ($\sin 2\phi$, $\cos 2\phi$) in vowel space. Whether the softness and harshness perception mentioned contributes to some consonant perceptions will be investigated next year in relation to the next phase of our research. However, within the vowel space concept we feel the question of the $\sin 3\phi$, $\cos 3\phi$ terms is important. To test the existence of these terms we turned up and down simultaneously triads of filters with centers 120 degrees apart on the vowel circle. We found again that when coordinated carefully the net effect seemed to be an increase of loudness and no change in vowel quality. We conclude that the higher terms discussed here are missing in vowel perception and the vowel space is, to a good approximation, three dimensional. The statement may not be true for fricatives and consonants.

19. From tristimulus theory it is expected that there should be an optimum setting for three narrow filters for purposes of speech intelligibility. For the recorded speeches of President Kennedy we found that the best setting corresponded approximately to 500, 1250, 3125. Note that this set roughly represents the maxima of tristimulus curves. The addition of a fourth filter, 315, improved voice quality (which is beyond color analogy). We found that the optimum setting was different for different speakers, but the variation was not large. The areas 450, 1400, and 3000 probably form an optimum setting.

20. To test the second order effect $\alpha_o' = \alpha_o(1 - \sigma^2/\Sigma^2)^{1/2}$ is not easy; firstly, the effect is not large; secondly, it is always simultaneously present with a change in vowel quality so that one must be able to distinguish intensity increase from a change in vowel quality. To the best of our judgment we observed the fact that if two complementary vowels are switched back and forth, say, with an interval of two seconds, the one newly turned on sounds louder. The loudness difference depends on the relative saturation of the two vowels. For small values of relative saturation the formula $\alpha_o' = \alpha_o(1 - \sigma^2/\Sigma^2)^{1/2}$, where σ and Σ are relative and maximum saturations, respectively, seemed to be satisfied. For large values of σ the formula breaks down. Since this experiment might depend critically on the switching time and switching transient we refrain from drawing more definite conclusions at the present moment. The matter will be investigated further in the future.

21. The analogue of the Bezold-Brucke phenomenon of color vision (the shifts of hues under large changes in intensity) was also investigated. It was found that there are similar nonlinear saturation phenomena in vowel percpetion. These work essentially in the same way as in color; namely, when the intensity is increased there exists a systematic shift toward 800 (o) and 2500 (e). In other words there are four stable points on the vowel circle--800, 2500 and 400,1250--such that shifts occur toward 800, 2500 when intensity is increased; whereas they tend toward 400, 1250 when intensity decreased. These observations will be refined in the future.

In conclusion, the vowel space previously devised seems to accommodate all known and newly uncovered effects to a good approximation. For purposes of intelligibility no further structure seems necessary although for speaker recognition one or two more response functions might possibly be required.

V. DISCUSSION AND COMPARISON WITH OTHER APPROACHES

Certain crude analogies between vowel and color were noted by early workers. For example, the formant theory of vowels resembles color theory in its three formants, and the vocoder approach uses a representation which may be said to be analogous to color. In fact, such homological concepts play some part in the present day research on speech.

As a result of our theoretical ideas (evolutionary theory of perception) and the experiments we have so far performed, we have come to believe that the existing formant and vocoder concepts are too vague and restrictive to be efficient guides for further progress in a field as sophisticated and exacting as the computer recognition of speech.

What seems to be emerging from our work is a consistent and comprehensive theory of speech in general and of vowels in particular. In order to compare our ideas with the current directions to speech we shall here summarize briefly some of the existing approaches. A few words of comparison will be added for each approach.

A. FORMANT THEORY AND FORMANT TRACKING APPROACH

This theory deals with the individual sharp energy peaks in the sound distribution. Proponents of this theory believe that the problem of vowel recognition can be solved by tracking the formants and devising some sort of a recognition algorithm. However, formants change their positions from person to person so widely that the vowel representation in terms of them become fuzzy and spread out. In fact, sometimes we may have ten or more equally sharp peaks to choose from, whereas in other circumstances it may be impossible to identify a desired formant. The problem of formant is a very much confused issue. For example,

Sovijärvi reported finding eleven fixed and seven variable formants. Fukumura tried to define average formants by summing spectral peaks of a number of subjects graphically.

We believe such attempts do not go to the root of the matter. Our perceptual approach predicts theoretically (and this may now be considered proven experimentally) that the wider ranges of frequency distributions and their over-all balance is important. Note in particular that the perceptual effect of a given frequency is not proportional to its physical intensity but to the logarithms of the physical intensity. The logarithmic weight reduces the efficiency of high energy peaks and increases the efficiency of lower areas of distribution. This makes the exaggerated expectations of formant theories rather faulty and inaccurate. Furthermore, formant approach does not give any clue towards the various adaptation and transformation phenomena that our theory naturally introduces, nor does it shed any light on the problem of consonants, plosives, and time-context. The most useful aspect of formant theory has been the three formant idea, which by careful choice of their positions and a clever, albeit ambiguous, identification has a resemblance to our tristimulus functions. Our theory shows the ambiguous nature of formant representation by the existence of vowel transformations. (for example, the perception of (i) in Land analogy, although no formant exists to produce this sound physically.) It also explains why the three formant aspect is useful as a crude approximation to vowel perception.

B. VOCODER (VOICE CODER) APPROACH

Vocoder was introduced originally (1930's) as a speech synthesis approach by Dudley. It was the result of ten years of development. The basic principle of vocoder is that it detects human speech, then transmits a description of this speech in a time-frequency code rather than the original signal itself. This code depends on a frequency analysis and aims at a bandwidth compression. Nowadays the term "vocoder" is used

in a generic sense for any effort aimed at frequency analysis and bandwidth compression.

Note, however, that vocoder approach by itself is not a theory and it must make use of some theory in order to achieve its goal. But it can be used as a research tool to arrive at a workable theory or to test the relative merits of various theories. In fact, some of the inadequacies of the formant theories were uncovered by vocoder people. Some new developments such as the pattern playback, filtered speech, and the discovery of k-t-p effect also resulted indirectly from the vocoder efforts. On the whole vocoder approach failed to suggest a new theory or point clearly to some definite direction. This is probably due to the strong preconceived assumption underlying the whole approach; namely, the assumption that the physical frequency composition of a sound distribution is also its perceptual determinant. For an engineer, this is a very natural assumption. However, the perception theory shows that the human ear goes beyond the merely absolute physical quantities and operates on the basis of transformations, relative values, and the extraction of invariants of a physical distribution.

C. SEGMENTATION APPROACH IN SPEECH RECOGNITION

This approach deals with the time behavior of speech within individual, separately pronounced words. The underlying assumption seems to be that phonemes are separated by certain detectable physical characteristics (for example, a dip in energy content). It is hoped that by detecting these a word can be segmented to facilitate recognition. On the basis of this, some decision algorithms are devised and computer programs written. Note that in this approach the frequency analysis part requires a separate theory and for this the formant theory is usually made use of. The segmentation approach may be viewed as a beginning of a theory of time dependence of speech. However, at the present time it is in a primitive stage and does not pay attention to perceptual effects

in time variable, nor to the interrelationships (physical or perceptual) of time with the frequency distribution.

In our theory speech is a pattern $f(\nu,t)$ in time-frequency space. Both physical and perceptual effects are to be considered in this over-all space. The frequency part is already investigated to some satisfactory stage. During the next six months or so we shall be investigating the time-dependence. According to our theory, the time-dependence manifests itself as transitions between various relevant points within the vowel space. Note that these transitions are not necessarily identical to physical variations which induce them. By way of an example, let a given transition represent the syllable (da). We expect that many different physical situations will result in the same perception (homomorphism) and the same physical stimulus will result in different perceptions under different adaptation and contextual conditions (transformation). We shall be trying to determine these transformations.

D. SPEECH PRODUCTION, VOCAL TRACT, AND MOTOR APPROACHES

In this approach attention is focused on the production mechanism. The vocal tract is studied and analyzed in great detail. Continuous or lumped circuit representations are devised and differential equations written down. To our theory this activity is highly significant because it helps to define the sound environment in which the perception of the ear evolves. Together with the natural sounds such as clicks and winds, etc., the vocally produced sounds provide an over-all sound environment (similar to light in our surroundings). In this environment the ear evolves some perceptual ability. This in turn influences the vocal tract. In this way a nonlinear interaction takes place and essentially both the vocal tract and the ear try to adapt to each other. However, it would be wrong to assume that vocal tract determines completely the perception of speech. For example, whispered speech is perfectly intelligible; whereas the vocal tract produces also voices and whistles. Furthermore,

a given sound produced by the vocal tract does not always give rise to the same perception. For example, if in the word "kiss" the "i" is replaced physically by "a", the perception of "k" will change into "t". If it is replaced by "u", then "k" will sound like "p". These and similar examples show that production and recognition (perception) are different, albeit related, problems. In our theory they are separated so that the production process and the properties of the vocal tract are used to define the environment. In this connection we may mention the advance production approach brought into the formulation of the problem of a formant. K. N. Stevens, after a careful analysis, defined the formant as a proper vibration--with a center frequency and half width--of the vocal tract. The formants then become simply analogous to the spectral lines of atoms in light production.* Indeed, the atomic spectral lines have also a center frequency and a Lorentzian half-width. Thus in our vowel theory the individual formants should map to the most saturated outer areas of the vowel circle, that is, they must define the boundaries of the vowel cone.

*Although our speech theory can be presented and formulated independently of color vision, we refer very often to a color-speech analogy because of the existing convenient terminology in color. Also color analogy has definite appeal to physicists and physically trained scientists.

VI. CONCLUSIONS AND POSSIBLE APPLICATIONS

In conclusion, our theory does not contradict or destroy the above approaches. Rather, it is a general theory which is capable of incorporating the useful aspects of each of them. It promises a consistent basis for the whole speech process including transformations and suggests valuable applications toward the eventual solutions of long persistent problems in speech recognition and speech synthesis. In achieving this consistency and generality, however, a number of old cherished ideas have to be abandoned and a rather abstract framework be adopted. In order to emphasize the points of divergence we may now summarize the most relevant experiments and their conceptual implications.

1. It has been demonstrated that the whispered vowels organize into a circular arrangement in the order of u, o, a, ae, e, i, and u with gradiations in between. The vowel ɝ (a sound between i in bird and u in hue) falls at the center of the circle (Figure 7). This can be unambiguously displayed on the oscilloscope and implies a structural analogy between the perception of vowels by the ear and the perception of color by the eye.

2. It has been demonstrated that the mixture of two vowels falls within the vowel circle on the line joining them. More specifically, mixture is represented as the "centeroid" of the mixed vowels (for example, when u and a are mixed o is produced). This shows that as far as whispered vowels are concerned the ear does not work analytically but works synthetically as the eye does. According to a long-held view the ear is supposed to analyze the two vowels instead of making a new synthetic vowel sound out of them. This view can thus be abandoned.

3. Now, if we record a running speech from a radio announcer and filter this speech through an (u) filter (or simply add whispered (u)) according to the above, we now expect all the (a)s of the speaker to turn

into (o)s. Yet this does not happen. Speech is almost as natural as the original. This shows that simplistic theories based on physical variables alone will lead to contradictions; we need the perceptual transformations required by the evolutionary theory.

4. An extreme form of transformation is the perception of (i) in two-vowel perception. Although physically there is no energy in the regions corresponding to (i), the ear extracts an (i) by a transformation (Figures 13 and 14). This shows that the usual formant theories which depend on the physical energy peaks are wholly untenable. Therefore, they can also be abandoned.

Lest we give the impression of following a naive analogy between color and speech, it is desirable to emphasize that this is not so. The theory also shows that there are certain areas where no analogy whatever can be expected. For example, the eye covers less than one octave in the visible range, and therefore no analogue of harmonic and voice quality can exist in the eye. In this respect the ear is richer than the eye. On the other hand, the sound waves are scalar functions and cannot be polarized. Thus no ear can have the analogue of the polarization perception of the eye of a bee. Furthermore, by evolutionary necessity due to light and sound environment, the eye specialized in spacial variations of color (patterns), whereas the ear specialized in temporal variations (speech and music). Therefore, the analogy is only expected in the strict sense between nonvarying colors and whispered vowels. However, the theory goes one step further: a set of analogies is still expected between time-varying whispers (as in whispered speech) and space-varying colors (as in a painting or design). Such analogies will be considered in the near future. The study of totally nonanalogous aspects, such as the voice quality, music, and residue, must proceed strictly from the first principles.

Reviewing the whole approach we seem to possess for the first time a consistent theory of vowel perception. The most valuable

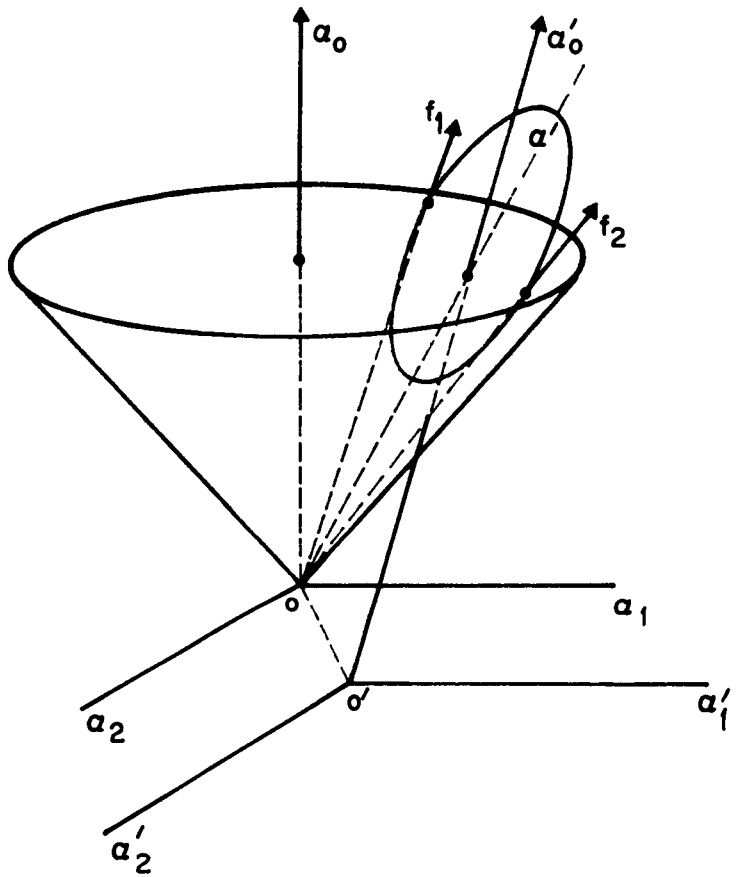


FIGURE 14 The two-filter experiment is explained by the transformation. The space spanned by the two vectors f_1, f_2 is practically a plane. Yet after the transformation a circular (more properly elliptical) arrangement around the new a'_0 axis is possible. Thus all vowel hues are obtainable by only two vowels. However, saturations will now depend on the vowel.

applications will probably be in speech analysis and synthesis, especially with regard to speech recognition. In speech recognition the translation of the psychophysical variables and their transformation into hardware now seems inevitable. Yet for this, more work is necessary. Vowels are a basic but relatively small part of the whole speech phenomenon. There are the consonants, glides, fricatives, and plosives. There is the persistent question of pitch, timbre, and residue. Lastly, there is the problem of context. One must study their perceptions and transformation properties before attempting ambitious recognition devices.

One particular finding, namely, the "voiced ($\mathcal{F}^{\nu}$)", seems to have immediate application in terms of speaker identification and natural speech. This point, in fact, is in the direction of "voice print" and voice identification research extensively studied elsewhere. Another finding, the two-vowel perception, offers bandwidth reduction possibilities over and above those presently possible.

Application of the developed device to teach the deaf how to speak offers considerable hope. Already the device is adequate for vowels in this respect. In the future, when the problems pertaining to the consonants and their transformations are also clarified, we expect that a valuable tool will be at hand for a fresh attack in this urgently needed area.

Compared to color and image making the state of the art in sound recording and reproduction is quite primitive. We have to record every vibration of sound as a primitive caveman trying to paint a tree had to rub the green leaves on the wall (advanced cavemen knew how to mix colors). Today an endless variety of colors can be synthesized out of only three suitably chosen primaries. Color photography, color movies, color television, textile, and other industries depend heavily on the knowledge of color perception. In sound we are heading toward a comparable science of speech. Such hitherto difficult problems as voice identity, bandwidth compression with natural tone, discrete mapping of phonemes, topological

voice encoding, speech recognition independent of speaker, speech synthesis, voice typewriter, and electronic machines taking spoken commands from a human operator will probably be solved when a comprehensive theory of speech is formulated. Considering the total investment of this country in the field of communication (probably no less than $100 billion), and considering the savings already in sight along perceptual lines, we are hoping that research in this direction will be pursued and encouraged by the scientific community. Speech as a tool of symbolic communication is the most important development in the evolution of man, and it distinguishes man from other animals. Any basic understanding, however small, into the nature of speech is of immense value to man.

In summary, we may say that the experiments so far have fully supported our original ideas on the nature of speech perception. As a consequence we now have the beginning of a general theory of speech. If further work substantiates the new direction, we may not be too far away from a fresh, all-out effort to tackle the residual and persistent problems of speech recognition. The general philosophy behind this mode of attack is evidently within the long anticipated theory of invariants. We are particularly encouraged by the fact that the fundamental theories of groups and invariants, as studied by mathematicians and modern physicists, find their way into the very workings of our perceptual makeup. The implementation of the theory of invariants into recognition devices of speech and pattern appears now both desirable and inevitable.

VII. REFERENCES

1. H. Yilmaz, "Color Vision and a New Approach to General Perception" Biological Prototypes and Synthetic Systems, Volume I (Plenum Press, 1962), pp. 126-141.

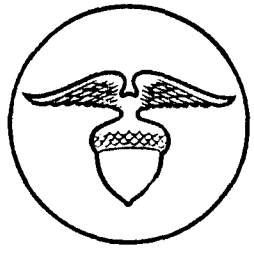
2. H. Yilmaz, "On Speech Perception," Report No. 42, Institute for Perception Research, Eindhoven, Holland (1964).

3. K. N. Stevens and A. S. House, "An Acoustical Theory of Vowel Production and Some of Its Implications," J. Speech and Hearing Research, Vol. 4, No. 4, 303-320 (1961).

4. S. S. Stevens, "On the Psychophysical Law," Psychol. Rev. 64, 153-181 (1957).

5. T. Chiba and M. Kajiyama, The Vowel: Its Nature and Structure (Phonetic Society of Japan, Tokyo, 1958).

6. G. Fant, Acoustic Theory of Speech Production (Mouton and Co., 1960).



CAMBRIDGE • WASHINGTON • EDINBURGH • CHICAGO
SANTA MONICA • MEXICO CITY • LONDON • ZÜRICH
BRUSSELS • SAN FRANCISCO • NEW YORK • TORONTO